DOCUMENT RESUME

ED 223 644 TM 820 719

AUTHOR Petersen, Nancy S.; And Others
TITLE Using Item Response Theory to Equate Scholastic Aptitude Test Scores.
PUB DATE 82
NOTE 65p.; Small print contained in many tables. Paper presented at the Annual Meeting of the American Psychological Association (Washington, DC, 1982).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *College Entrance Examinations; *Equated Scores; *Latent Trait Theory; *Research Methodology; Secondary Education; *Statistical Analysis; Test Theory; Weighted Scores
IDENTIFIERS Linear Equating Method; *Scholastic Aptitude Test

ABSTRACT

In January 1982, the College Board and Educational Testing Service implemented a technical change in the procedures used to equate scores on the Scholastic Aptitude Test (SAT). For previous editions of the SAT, a linear equating procedure was used to establish the comparability of scores on different editions. Beginning in January 1982, this equating procedure was replaced by an equating procedure based on item response theory (IRT, also known as item characteristic curve theory and latent trait theory). This change in equating procedures was preceded by a series of research studies, which indicated that the change was technically feasible and would result in improved maintenance of the score scale. Described is the long-range research and development program that has resulted in the change in equating procedures and promises further changes in the future. The results of this extensive research and development effort show that the new equating procedure will improve the comparability of scores on different editions of the SAT and the extent to which the SAT score scale maintains a constant meaning over time. Current directions of research on IRT equating of the SAT are also discussed. (Author/PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Using Item Response Theory to Equate Scholastic Aptitude Test Scores[1,2]

Nancy S. Petersen
Linda L. Cook
Gary L. Marco

Educational Testing Service

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

L. L. Cook

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

[1] A paper presented at the annual meeting of APA, Washington DC, 1982.

[2] Draft copy. Not for circulation without the express permission of the authors.

Introduction

In January 1982 the College Board and Educational Testing Service implemented a technical change in the procedures used to equate scores on the Scholastic Aptitude Test (SAT), which is offered by the College Board Admissions Testing Program. For previous editions of the SAT, a linear equating procedure was used to establish the comparability of scores on different editions. Beginning in January 1982, this equating procedure was replaced by an equating procedure based on item response theory (IRT, also known as item characteristic curve theory and latent trait theory). This change in equating procedures was preceded by a series of research studies, which indicated that the change was technically feasible and would result in improved maintenance of the score scale.

The purpose of this paper is not so much to justify the use of the new equating procedure for the SAT as to describe the long-range research and development program that has resulted in the change in equating procedures and promises further changes in the future. This paper demonstrates how a well-planned program of research and development can lead to operational improvements in testing programs.

Score equating has two fundamental goals: (1) fairness to the test-takers taking different editions of the test and (2) preservation of the meaning of the reported scores. Ideally, it should be a matter of indifference to test-takers at any ability level which edition (form) of the test they take; scaled scores on any two editions of the test should indicate the same levels of ability. In practice, these goals can never be perfectly achieved, but they can be approximated very closely - more closely by some equating methods than by others.

The first section of this paper describes the IRT equating procedure implemented in January 1982 for equating SAT scores. The decision to change equating procedures for SAT scores was preceded by a series of research studies

on the effects of using different equating methods with SAT score data. These studies are described in the second section of this paper. The first studies in this series were intended to compare the accuracy of various equating methods; the last two studies investigated the feasibility of IRT equating and the probable effects of a change to IRT equating. The final section of this paper describes research currently in progress and the possible changes this research could lead to in the future.

Item Response Theory Equating

Item response theory (IRT) assumes that there is a mathematical function which relates the probability of a correct response on an item to an examinee's ability. (See Lord, 1980, for a detailed discussion.) Many different mathematical models of this functional relationship are possible. The model that is used for the operational equating of the SAT and that was used for the studies carried out prior to implementation of IRT equating for the SAT is the three-parameter logistic model. In this model, where θ represents an examinee's ability, the probability of a correct response to item i, $P_i(\theta)$, is

$$P_i(\theta) = c_i + \frac{1-c_i}{1+e^{-1.7a_i(\theta-b_i)}} ,$$

where a_i, b_i, and c_i are three parameters describing the item. These parameters have specific interpretations: b_i is the point on the θ metric at the inflection point of $P_i(\theta)$ and is interpreted as the item difficulty; a_i is proportional to the slope of $P_i(\theta)$ at the point of inflection and represents the item discrimination; and c_i is the lower asymptote of $P_i(\theta)$ and represents a pseudo-guessing parameter.

The item parameters and examinee abilities required for IRT equating of SAT scores are estimated (calibrated) using the program LOGIST (Wood and Lord, 1976;

Wood, et al., 1976). The estimates are obtained by a (modified) maximum likelihood procedure with special procedures for the treatment of omitted items (see Lord, 1974). LOGIST requires as input the responses to a set of items from a group of examinees, coded to reflect items answered correctly, incorrectly, omitted, and not reached. In addition, the user may specify certain restrictions on the data and parameters in order to speed convergence of the iterative procedure.

LOGIST produces as output estimates of the a, b, and c for each item, and θ for each examinee. The metric, chosen arbitrarily for the θ (and b) scale, is such that the distribution of estimates of θ has mean zero and standard deviation one. If two separate LOGIST runs are made for the same items, but different groups of examinees, the resulting parameter estimates will be on different scales. IRT equating requires that parameter estimates for different editions of a test that are to be equated be on the same scale. Several methods for placing parameter estimates on the same scale were investigated prior to implementation of IRT equating for the SAT.

Once parameter estimates are obtained that are on the same scale, there are a number of equating techniques possible when using IRT; only true formula score equating (Lord, Chapter 13, 1980) was investigated and subsequently implemented for the SAT. The expected value of an examinee's observed formula score is defined as his or her true formula score. For the true formula score, ξ, we have

$$\xi = \sum_{i=1}^{n} \left[\frac{(k_i+1)}{k_i} P_i(\theta) - \frac{1}{k_i} \right]$$

where n is the number of items in the test and (k_i+1) is the number of choices for item i. If we have two tests measuring the same ability θ, then true formula scores ξ and η from the two tests are related by the equations

$$\xi = \sum_{i=1}^{n} \left[\frac{(k_i+1)}{k_i} P_i(\theta) - \frac{1}{k_i} \right]$$
$$\eta = \sum_{j=1}^{m} \left[\frac{(k_j+1)}{k_j} P_j(\theta) - \frac{1}{k_j} \right] \qquad (1)$$

Clearly, for a particular θ corresponding true scores ξ and η have identical meaning. They are said to be equated.

In practice, true score equating is carried out by substituting estimated parameters into equations (1). Paired values of ξ and η are then computed for a series of arbitrary values of θ. Since we cannot know an examinee's true formula score, we act as if relationships (1) apply to an examinee's observed formula score. That is, when this computation has been made for two editions of the SAT, the two expected formula scores corresponding to a given value of θ are considered equivalent and assigned the same scaled score.

Review of Research

Test Design

The first application of item response theory (IRT) to the SAT, interestingly enough, had nothing to do with equating. Rather it dealt with test design - developing new statistical specifications for the SAT. Lord (1977) discussed how a procedure based on IRT can be used to redesign a test - the SAT-Verbal in his example. He showed the effects on relative efficiency of modifying SAT-Verbal hypothetically in several ways. This procedure was used operationally to develop new statistical specifications for a shorter, slightly easier SAT, introduced in the fall of 1[illegible] an SAT that was more appropriate to the average test-taker but still discriminated about as well among high-ability test-takers as previous editions of the SAT.

IRT was also applied to evaluating a multilevel SAT-Mathematical and to pre-equating a miniature SAT-Verbal (Marco, 1977). This early work stimulated an

interest in the use of score equating procedures based on IRT, but researchers realized that, because of the strong assumptions of IRT, a long-range program of research was needed to evaluate the appropriateness of IRT equating for the SAT. It was clear from the very beginning, however, that IRT equating offered new hope for equating test scores when test editions and equating samples varied systematically. In particular, IRT equating filled a gap that had existed for years, namely, curvilinear true score equating. The only known procedures for curvilinear equating were intended for use with observed scores from random samples of test-takers (see Angoff, 1971).

Linear true score equating procedures did exist (Levine, 1957), but these could not take account of curvilinear trends that result at the ends of the test score range when test editions are not parellel or samples differ somewhat in ability.

Comparison of Equating Models

A large-scale simulation study based on real data was the first step on the way to using IRT equating for the SAT. The study was reported in two parts. The first part (Marco, Petersen, & Stewart, 1979) was a comparison of several linear equating methods, two equipercentile methods, and two IRT methods based on the one-parameter and three-parameter logistic test models. The second part (Petersen, Marco, & Stewart, 1981) was a comparison of linear methods and one equipercentile method. The purpose of the study was to find out to what extent, if at all, and under what circumstances newly developed equating methods were superior to conventional equating methods. Of particular interest were the IRT methods, which had only recently come into practical use.

All of the equating procedures in this study used anchor-test data, the kind of data routinely used for equating the SAT. An anchor-test design requires

administering one edition of a total test to one group of examinees, a second edition to a second group of examinees, and a common anchor test to both groups. The anchor test can be internal or external to the test to be equated. The anchor test is used to adjust for differences in ability between the two groups in order to reduce equating bias.

Linear, equipercentile, and IRT equating models (referred to as ICC or item characteristic curve models in the report) were studied. Linear equating transforms scores from the test to be equated to scores from the other test that are the same number of standard deviations from the mean in some group of examinees. Equipercentile equating transforms scores from the test to be equated to scores from the other test that have the same percentile ranks in some group of examinees. Item response theory (IRT) equating transforms scores from the test to be equated to scores on the other test that correspond to the same estimated levels of the latent trait as those on the test to be equated.

Linear and equipercentile equating methods using anchor-test data can be further classified as to whether equating is done directly or indirectly by frequency estimation. In direct equipercentile equating, only the marginal distributions of the total tests and anchor test are used; in indirect equipercentile equating the bivariate distributions of each of the total tests and the anchor test are used (see Angoff, 1971). Two methods of IRT equating were used, one based on the one-parameter logistic test model and one based on the three-parameter logistic test model. The one-parameter model was included in the study to allow a comparison of Rasch-type equating with equating derived from the three-parameter model. The computer program LOGIST was used to estimate the item parameters and examinee ability parameters for each model. True formula score equating procedures were used for both the one- and three-parameter logistic models.

A total of 22 different equating models were used in the case of an external anchor test and 20 in the case of an internal anchor test (see Table 1 in Petersen, Marco, & Stewart, 1981). Included among these were procedures described by Pothoff (1966) and Angoff (1971).

Six base samples were used for the study, three each from the April 1975 and November 1975 SAT administrations. Each base sample consisted of 4,731 cases, from which three nonoverlapping samples of 1,577 cases each were created in three different ways. Three random samples, three similar samples, and three dissimilar samples were created for each base sample. Thus, a total of 54 subsamples of 1,577 cases each, nine for each of the base samples, were used in the study. In terms of mean scores on SAT-Verbal, the differences in ability represented in any given set of three samples ranged from 4 to 8 points for random samples, 13 to 24 points for similar samples, and 74 to 87 points for dissimilar samples.

Several scores calculated as part of the normal processing for SAT administrations were included in the study. Included also were scores computed on a number of tests constructed from the items in the base tests solely for use in this study. The general approach followed in constructing the special-purpose tests entailed (a) developing content and statistical specifications and (b) identifying sets of items in accordance with these specifications.

For any one of the three base samples for a given administration, scores were available on only one of the three external anchor tests (a non-operational section of verbal material given with the SAT) similar in content and difficulty to SAT-Verbal. Two tests administered with SAT-Verbal were also used as external anchor tests; SAT-Mathematical and the Test of Standard Written English (TSWE).

For each of the two administrations, three internal anchor tests for equating SAT-Verbal to itself were specially contructed for the study from the pool of 85 operational items in each edition. The internal anchor tests were constructed to be similar in content to SAT-Verbal but to vary with regard to each other in mean difficulty. Easy, medium, and hard anchor tests were constructed. A fourth anchor test was assembled solely from antonym and analogy items, two of the four item types used in SAT-Verbal. Reading items were excluded. Although it was dissimilar in content to the SAT-Verbal, it was made similar in difficulty.

Additional tests were constructed for equating a test to a different test rather than to itself. For each administration, an expanded item pool consisting of the 85 SAT-Verbal items, the 40 items in one of the verbal external anchor tests, and the 50 items in one form of the TSWE formed the basis for defining four sets of three total tests each. The tests in one set all contained the same number of items but varied in difficulty. The tests within each of the other three sets were systematically different in average difficulty, length, or content, but were similar in the other two respects. For each administration, a 20-item internal anchor test of medium difficulty similar in content to SAT-Verbal was constructed for use in equating tests that were similar in content. A 21-item internal anchor test, composed of vocabulary reading, and TSWE items was built to equate tests dissimilar in content. Exclusive of this anchor test, each of these tests contained only one of the three content categories (reading, vocabulary, or English).

The total-test anchor-test combinations used in the various equatings for each of the two administrations were classified into 11 categories, referred to here as test variations. The test variations studied are shown in Table 1.

TABLE 1

Test Variations Incorporated in Equating Design

Test variation	Comparison of total tests[a]			Anchor test in relation to total tests		
	Content	Difficulty	Length	Location	Content	Difficulty
1	Identical	Equal	Equal	External	Similar	Similar
2	Identical	Equal	Equal	External	Dissimilar	Similar
3	Identical	Equal	Equal	External	Dissimilar	Dissimilar
4	Identical	Equal	Equal	Internal	Similar	Similar
5	Identical	Equal	Equal	Internal	Similar	Dissimilar
6	Identical	Equal	Equal	Internal	Dissimilar	Similar
7	Similar	Similar	Unequal	Internal	Similar	Similar
8	Similar	Dissimilar	Equal	Internal	Similar	Similar to one total test
9	Similar	Dissimilar	Equal	Internal	Similar	Between total tests
10	Dissimilar	Similar	Equal	Internal	Dissimilar	Similar
11	Similar	Similar	Equal	Internal	Similar	Similar

[a]Test variations 1-6 involved equating a test to itself; hence, the content, difficulty, and length were exactly the same. Test variations 7-11 involved equating a test to a different test.

Two kinds of equating were conducted. In one case a single form of the operational portion of the SAT-Verbal was treated as if it represented two different forms, and it was equated to itself. This type of design was first used by Levine (1955). Six test variations (Variations 1-6) for this type of equating were defined in terms of the characteristics of the anchor test in relation to the total test with regard to location (external or internal), content (similar or dissimilar), and difficulty (similar or dissimilar).

In the other case, different total tests constructed from the item pool for the study were equated. Five test variations (Variations 7-11) for equating a test to a different test through an internal anchor test were defined by the relations between the total tests and the relations of the anchor test to the total tests. All test variations were involved in the comparisons of linear and equipercentile equating models. Only test variations 1, 4, 8, and 9 for related random samples and dissimilar samples were involved in the comparisons of curvilinear equating models because of the expenses associated with these equatings. The mean square difference, weighted by the total sample, was used to evaluate the effectiveness of the various models. This index is made up of two components: variance of the difference and mean difference squared (squared bias). The values were standardized so that the criterion standard deviation was 100 for the standard reference group.

Where a test was equated to itself, the criterion for the various equatings was the test score itself. The "new" and "old" editions were treated as different tests when in reality they were one and the same test. The ideal equating would reproduce the score on the old edition exactly.

Where a test was equated to a different test, the criterion was established by equating the total tests using data from all of the cases in the two base

samples from which the subsamples were selected. Since all 4,731 cases in each base sample had scores on the tests being experimentally equated, the scores could be linked directly without involving an anchor test. Two equating methods were used for these equatings: equipercentile equating of estimated true scores derived from the three-parameter logistic test model, "the ICC (item characteristic curve) equipercentile criterion", and equipercentile equating of observed scores, "the direct equipercentile criterion." Both criteria were used in the first part of the study, which involved the comparisons of the curvilinear methods used in the study. Only the ICC equipercentile criterion was used in the second part of the study, which did not include the frequency estimation and IRT equating methods.

Tables 2 and 3 give the summary results of the study. Table 2 is a summary of the results given in Marco, Petersen, and Stewart (1979); Table 3 is reproduced from the Petersen, Marco, and Stewart paper (1981). In the first table the various curvilinear methods were compared with the best linear method. Codes are used in the table to stand for the various equating models that were used (see the original reports for full descriptions). The curvilinear methods are represented by the following codes: EQ (equipercentile equating), FE-EQ (frequency estimation equating), ICC (three-parameter logistic model), and Rasch (one-parameter logistic model). The equipercentile model (EQ) is the only curvilinear model represented in Table 3; the remaining models are linear models.

In the tables the weighted mean square difference or total error was classified into one of five zones (A - E), representing different degrees of acceptability. Zone E was considered unacceptable; the other zones were acceptable. Zone A was based on what would be expected from random sampling when a test is equated through an appropriate anchor test. Two criteria were used for evaluating

Table 2

Total Weighted Mean Squared Error in Various Zones (A-F)[a]

for Experimental Curvilinear Equatings

Test Variation	Equating Model	Random Samples						Dissimilar Samples					
		No. of Equatings	A	B	C	D	E	No. of Equatings	A	B	C	D	E
1													
Equating SAT-V to Itself through an External Anchor Test Similar in Content and Difficulty	Best Lin.	2	2	-	-	-	-	2	2	-	-	-	-
	EQ	2	1	1	-	-	-	2	-	2	-	-	-
	FE-EQ	2	1	1	-	-	-	2	-	1	1	-	-
	ICC	2	1	1	-	-	-	2	1	1	-	-	-
	RASCH	2	2	-	-	-	-	2	2	-	-	-	-
4													
Equating SAT-V to Itself through an Internal Anchor Test Similar in Content and Difficulty	Best Lin.	2	2	-	-	-	-	2	2	-	-	-	-
	EQ	2	2	-	-	-	-	2	2	-	-	-	-
	FE-EQ	2	2	-	-	-	-	2	-	2	-	-	-
	ICC	2	2	-	-	-	-	2	2	-	-	-	-
	RASCH	2	2	-	-	-	-	2	2	-	-	-	-
5													
Equating SAT-V to Itself through an Internal Anchor Test Similar in Content but Dissimilar in Difficulty	Best Lin.	4	4	-	-	-	-	4	2	2	-	-	-
	EQ	4	3	1	-	-	-	4	3	1	-	-	-
	FE-EQ	4	3	1	-	-	-	4	1	3	-	-	-
	ICC	4	4	-	-	-	-	4	4	-	-	-	-
	RASCH	4	4	-	-	-	-	4	4	-	-	-	-
8													
Equating Different Tests Similar in Content but Somewhat Dissimilar in Difficulty through an Internal Anchor Test Similar in Content (ICC Equipercentile Criterion)	Best Lin.	4	-	-	-	-	4	4	-	-	-	-	4
	EQ	4	-	4	-	-	-	4	-	4	-	-	-
	FE-EQ	4	2	2	-	-	-	4	-	2	2	-	-
	ICC	4	2	2	-	-	-	4	3	1	-	-	-
	RASCH	4	-	1	2	1	-	4	-	1	2	1	-
8													
Equating Different Tests Similar in Content but Somewhat Dissimilar in Difficulty through an Internal Anchor Test Similar in Content (Direct Equipercentile Criterion)	Best Lin.	4	-	-	-	3	1	4	-	-	-	2	2
	EQ	4	4	-	-	-	-	4	1	3	-	-	-
	FE-EQ	4	4	-	-	-	-	4	1	2	1	-	-
	ICC	4	1	3	-	-	-	4	1	3	-	-	-
	RASCH	4	-	2	2	-	-	4	-	2	2	-	-

Table 2 (Continued)

Test Variation	Equating Model	Random Samples						Dissimilar Samples					
		No. of Equatings	A	B	C	D	E	No. of Equatings	A	B	C	D	E
9													
Equating Different Tests Similar in Content but Very Dissimilar in Difficulty through an Internal Anchor Test Similar in Content (ICC Equipercentile Criterion)	Best Lin.	2	-	-	-	-	2	2	-	-	-	-	2
	EQ	2	-	1	1	-	-	2	-	1	1	-	-
	FE-EQ	2	-	1	1	-	-	2	-	-	1	1	-
	ICC	2	1	1	-	-	-	2	-	2	-	-	-
	RASCH	2	-	-	-	-	2	2	-	-	-	1	1
9													
Equating Different Tests Similar in Content but Very Dissimilar in Difficulty through an Internal Anchor Test Similar in Content (Direct Equipercentile Criterion)	Best Lin.	2	-	-	-	-	2	2	-	-	-	-	2
	EQ	2	2	-	-	-	-	2	-	2	-	-	-
	FE-EQ	2	2	-	-	-	-	2	-	-	1	1	-
	ICC	2	-	-	1	-	-	2	-	2	-	-	-
	RASCH	2	-	-	-	2	-	2	-	-	1	1	-

[a]Zone A: Total Error $\leq$ 25

B: 25 < Total Error $\leq$ 100

C: 100 < Total Error $\leq$ 225

D: 225 < Total Error $\leq$ 400

E. 400 < Total Error

Table 3

Total Weighted Mean Squared Error in Various Zones (A-E)[a] for Experimental Linear Equatings

Test Variation	Model	Random Samples						Similar Samples						Dissimilar Samples					
		No. of Equatings	A	B	C	D	E	No. of Equatings	A	B	C	D	E	No. of Equatings	A	B	C	D	E
1																			
Equating SAT-V to Itself through an External Anchor Test Similar in Content and Difficulty	T1	18	16	2	-	-	-	18	4	13	1	-	-	18	1	8	6	2	1
	LER	18	16	2	-	-	-	18	13	5	-	-	-	18	9	7	2	-	-
	T2T	18	15	3	-	-	-	18	12	6	-	-	-	18	9	7	2	-	-
	LUR	18	16	2	-	-	-	18	13	5	-	-	-	18	9	7	2	-	-
	LCS	18	16	2	-	-	-	18	13	5	-	-	-	18	7	9	2	-	-
	PAC	18	7	10	1	-	-	18	-	-	-	3	15	18	-	-	-	1	17
	PB	18	17	1	-	-	-	18	15	3	-	-	-	18	5	9	3	1	-
	PD	18	16	2	-	-	-	18	13	3	-	-	-	18	8	9	1	-	-
2	EQ	18	9	9	-	-	-	18	6	12	-	-	-	18	8	8	1	1	-
Equating SAT-V to Itself through SAT-M	T1	18	17	1	-	-	-	18	1	8	7	2	-	18	-	-	5	2	11
	LER	18	17	1	-	-	-	18	11	7	-	-	-	18	4	9	4	1	-
	T2T	18	13	5	-	-	-	18	6	12	-	-	-	18	4	9	4	1	-
	LUR	18	17	1	-	-	-	18	8	10	-	-	-	18	4	9	3	1	1
	LCS	18	13	5	-	-	-	18	3	5	6	3	1	18	1	1	8	4	4
	PAC	18	6	10	2	-	-	18	-	-	-	1	17	18	-	-	-	1	17
	PB	18	[illegible]	4	1	-	-	18	-	6	8	2	2	18	1	-	2	4	11
	PD	[illegible]	[illegible]	1	-	-	-	18	11	6	1	-	-	18	3	8	6	1	-
3	EQ	18	5	13	-	-	-	[illegible]	3	12	3	-	-	18	-	11	6	1	-
Equating SAT-V to Itself through TSWE	T1	18	14	4	-	-	-	18	6	6	3	3	-	18	-	2	6	4	6
	LER	18	12	6	-	-	-	18	1	5	9	2	1	18	3	6	4	3	2
	T2T	18	11	7	-	-	-	18	1	4	10	2	1	18	2	3	5	3	4
	LUR	18	13	5	-	-	-	18	-	6	9	2	1	18	3	5	3	4	3
	LCS	18	14	4	-	-	-	18	3	1	9	2	3	18	1	5	3	4	5
	PAC	18	8	9	1	-	-	18	-	-	-	3	15	18	-	-	-	1	17
	PB	18	14	4	-	-	-	18	2	2	3	7	4	18	1	2	5	2	8
	PD	18	14	4	-	-	-	18	1	8	6	3	-	18	2	7	4	3	2
4	EQ	18	5	13	-	-	-	18	1	9	6	2	-	18	2	15	1	-	-
Equating SAT-V to Itself through an Internal Anchor Test Similar in Content and Difficulty	T1	18	18	-	-	-	-	18	16	2	-	-	-	18	8	8	2	-	-
	LER	18	18	-	-	-	-	18	18	-	-	-	-	18	17	1	-	-	-
	T2T	18	18	-	-	-	-	18	18	-	-	-	-	18	17	1	-	-	-
	LUR	18	18	-	-	-	-	18	18	-	-	-	-	18	17	1	-	-	-
	PAC	18	7	9	2	-	-	18	-	-	-	4	14	18	-	-	-	1	17
	PD	18	18	-	-	-	-	18	18	-	-	-	-	18	15	3	-	-	-
	EQ	18	16	2	-	-	-	18	13	5	-	-	-	18	9	9	-	-	-
5																			
Equating SAT-V to Itself through an Internal Anchor Test Similar in Content but Dissimilar in Difficulty	T1	34	34	2	-	-	-	36	20	16	-	-	-	36	3	11	14	5	3
	LER	36	34	2	-	-	-	36	20	15	1	-	-	36	3	11	9	8	5
	T2T	36	33	3	-	-	-	36	17	17	2	-	-	36	1	12	10	4	9
	LUR	36	33	3	-	-	-	36	20	16	-	-	-	36	3	10	11	9	3
	PAC	36	14	18	2	2	-	36	-	-	1	4	31	36	-	1	-	2	33
	PD	36	34	2	-	-	-	36	23	13	-	-	-	36	4	12	12	5	3
	EQ	36	32	4	-	-	-	36	20	16	-	-	-	36	12	21	3	-	-
6																			
Equating SAT-V to Itself through an Internal Vocabulary Test Similar in Difficulty	T1	18	18	-	-	-	-	18	15	3	-	-	-	18	3	10	5	-	-
	LER	18	18	-	-	-	-	18	18	-	-	-	-	18	11	7	-	-	-
	T2T	18	18	-	-	-	-	18	18	-	-	-	-	18	10	8	-	-	-
	LUR	18	18	-	-	-	-	18	18	-	-	-	-	18	12	6	-	-	-
	PAC	18	8	10	-	-	-	18	-	-	-	3	15	18	-	-	-	1	17
	PD	18	18	-	-	-	-	18	18	-	-	-	-	18	11	7	-	-	-
	EQ	18	15	3	-	-	-	18	15	3	-	-	-	18	5	13	-	-	-

Table 3 (cont'd)

Test Variation	Model	Random Samples						Similar Samples						Dissimilar Samples					
		No. of Equatings	A	B	C	D	E	No. of Equatings	A	B	C	D	E	No. of Equatings	A	B	C	D	E
7																			
Equating Different Tests Similar in Content and Difficulty but Unequal in Length through an Internal Anchor Test Similar in Content and Difficulty	T1	10	8	2	-	-	-	6	3	3	-	-	-	6	-	3	3	-	-
	LER	10	9	1	-	-	-	6	6	-	-	-	-	6	3	3	-	-	-
	T2T	10	8	2	-	-	-	6	5	1	-	-	-	6	5	2	-	-	-
	LUR	10	8	2	-	-	-	6	5	1	-	-	-	6	5	1	-	-	-
	PAC	10	-	1	-	3	6	6	-	2	1	1	2	6	-	-	-	-	6
	PD	10	9	1	-	-	-	6	4	2	-	-	-	6	-	6	-	-	-
	EQ	10	6	4	-	-	-	6	4	2	-	-	-	6	-	5	1	-	-
8																			
Equating Different Tests Similar in Content but Somewhat Dissimilar in Difficulty through an Internal Anchor Test Similar in Content	T1	6	-	-	-	-	6	4	-	-	-	-	4	4	-	-	-	-	4
	LER	6	-	-	-	-	6	4	-	-	-	-	4	4	-	-	-	-	4
	T2T	6	-	-	-	-	6	4	-	-	-	-	4	4	-	-	-	-	4
	LUR	6	-	-	-	-	6	4	-	-	-	-	4	4	-	-	-	-	4
	PAC	6	-	-	-	-	6	4	-	-	-	-	4	4	-	-	-	-	4
	PD	6	-	-	-	-	6	4	-	-	-	-	4	4	-	-	-	-	4
	EQ	6	-	6	-	-	-	4	-	4	-	-	-	4	-	4	-	-	-
9																			
Equating Different Tests Similar in Content but Very Dissimilar in Difficulty through an Internal Anchor Test Similar in Content	T1	4	-	-	-	-	4	2	-	-	-	-	2	2	-	-	-	-	2
	LER	4	-	-	-	-	4	2	-	-	-	-	2	2	-	-	-	-	2
	T2T	4	-	-	-	-	4	2	-	-	-	-	2	2	-	-	-	-	2
	LUR	4	-	-	-	-	4	2	-	-	-	-	2	2	-	-	-	-	2
	PAC	4	-	-	-	-	4	2	-	-	-	-	2	2	-	-	-	-	2
	PD	4	-	-	-	-	4	2	-	-	-	-	2	2	-	-	-	-	2
	EQ	4	-	2	2	-	-	2	-	1	1	-	-	2	-	1	1	-	-
10																			
Equating Different Tests Dissimilar in Content but Similar in Difficulty through an Internal Anchor Test Dissimilar in Content but Similar in Difficulty	T1	10	6	4	-	-	-	6	1	5	-	-	-	6	2	3	1	-	-
	LER	10	5	5	-	-	-	6	4	2	-	-	-	6	2	4	-	-	-
	T2T	10	4	6	-	-	-	6	4	2	-	-	-	6	2	4	-	-	-
	LUR	10	6	4	-	-	-	6	3	3	-	-	-	6	2	4	-	-	-
	PAC	10	1	8	-	1	-	6	-	1	-	3	2	6	-	-	-	-	6
	PD	10	6	4	-	-	-	6	3	3	-	-	-	6	3	3	-	-	-
	EQ	10	3	7	-	-	-	6	3	3	-	-	-	6	2	4	-	-	-
11																			
Equating Different Tests Similar in Content and Difficulty through an Internal Anchor Test Similar in Content and Difficulty	T1	10	10	-	-	-	-	6	4	2	-	-	-	6	-	3	2	1	-
	LER	10	9	1	-	-	-	6	5	1	-	-	-	6	4	2	-	-	-
	T2T	10	9	1	-	-	-	6	5	1	-	-	-	6	4	2	-	-	-
	LUR	10	10	-	-	-	-	6	5	1	-	-	-	6	4	2	-	-	-
	PAC	10	4	3	3	-	-	6	-	-	-	4	2	6	-	-	-	-	6
	PD	10	10	-	-	-	-	6	4	2	-	-	-	6	1	4	1	-	-
	EQ	10	8	2	-	-	-	6	2	4	-	-	-	6	1	4	1	-	-

*Zone A: Total Error $\leq$ 25

B: 25 < Total Error $\leq$ 100

C: 100 < Total Error $\leq$ 225

D: 225 < Total Error $\leq$ 400

E: 400 < Total Error

the models represented in Table 2; only the ICC equipercentile criterion was used for evaluating the models represented in Table 3.

Based on the results presented in these tables and other data from the two reports on this study, the following conclusions may be drawn:

1. When a test is equated to itself (Test Variations 1-6) or to a similar test (Test Variation 11):
   a. The equating error is small for almost all of the equating models when random samples are used;
   b. When dissimilar samples are used, however, the total error is smaller for those models that require error variance estimates.
   c. Except for frequency estimation equipercentile equating, curvilinear equating is superior to linear equating when a nonlinear relationship, due to differences in difficulty, exists between the anchor test and the total tests (Test Variations 3 and 5) and dissimilar samples are used.
   d. Differences in difficulty between the anchor test and total tests lead to greater equating error than moderate differences in content between the anchor test and total tests.
   e. The one- (Rasch) and three-parameter (ICC) models perform similarly. (The models were not evaluated, however, for Test Variation 11.)
   f. Some criterion bias favoring the one-parameter logistic model (Rasch) seems to exist when a test is equated to itself.
2. When a test is equated to a test differing in content, difficulty, or length (Test Variations 7-10):
   a. Linear models perform poorly when a nonlinear relationship, due to differences in difficulty, exists between the total tests to be equated.

b. Differences in difficulty between the two tests to be equated lead to greater equating error than moderate differences in content.

c. The one-parameter logistic model (Rasch) yields a larger amount of equating error than any of the other curvilinear models in most situations in which the tests to be equated differ in difficulty.

d. Some criterion bias favoring either the direct equipercentile model (EQ) or the three-parameter logistic model (ICC) seems to exist.

3. The most consistently accurate equating model over a wide range of conditions is the three-parameter logistic model (ICC).

This study was important in that it demonstrated for the first time that IRT equating based on the three-parameter logistic model is an acceptable equating procedure. But questions remained. Because of the limited number of comparisons among curvilinear models and the possibility of criterion bias, no definitve conclusions could be reached regarding the superiority of IRT equating. Furthermore, the study had used an SAT-Verbal item pool. SAT-Mathematical items also needed to be investigated.

SAT Scale Drift

The next study (Petersen, Cook, and Stocking, 1981) was designed to determine which of several equating procedures would best preserve the meaning of SAT-Verbal and SAT-Mathematical scores over time. Scale drift contributing to lack of score comparability over time, will occur if the result of equating, say, edition A (the new edition) to edition D (the old edition) directly is not the same as that obtained by equating edition A to edition D through intervening editions B and C.

Scale drift for the verbal and mathematical sections of the SAT was investigated using linear, equipercentile, and item response theory (IRT) equating methods. Three linear models, namely, Tucker, Levine Equally Reliable, and Levine Unequally Reliable,

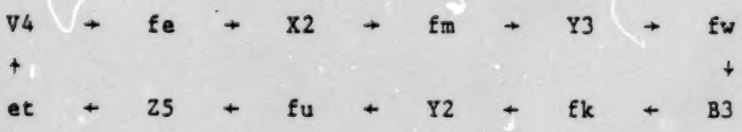
were used. Three IRT calibration designs, used to place parameter estimates on the same scale prior to equating, were investigated. The IRT designs are referred to as (1) concurrent, (2) partial pre-calibration with fixed b's, and (3) partial pre-calibration with equated b's. Figure 1 illustrates the chain of six verbal equatings and the chain of six mathematical equatings that were used for the study. Upper case letters indicate SAT editions and lower case letters indicate anchor (equating) tests. It was possible to use each chain to assess scale drift in the following manner. Edition V4, which had previously been put on scale, was equated to itself through the five intervening editions X2, Y3, B3, Y2, and Z5. This circularity provided a criterion for accuracy; the raw-to-scaled score conversion for edition V4 should be exactly the same after the six equatings as before. Any discrepancy between the transformation obtained from the circular chain of equatings and the initial V4 scale (the criterion scale) was attributed to scale drift.

A sample of approximately 2670 cases was selected for each pairing of an SAT edition and an anchor test shown in Figure 1. The samples were randomly selected from data obtained at the respective test administrations. The test administration dates and summary statistics for each verbal and mathematical sample are given in Table 4.

For the IRT concurrent calibration design, each pair of verbal or mathematical test editions plus associated anchor test was calibrated in a single LOGIST run (see Figure 2). This results in item parameters on a common scale for each pair and allows direct equating of the two editions.

For the two partial pre-calibration designs (fixed b's and equated b's) LOGIST runs were made separately for each test edition and associated anchor test (see Figure 3). For the fixed b's procedure, items were calibrated sequentially,

Chain of Six Verbal Equatings



Chain of Six Mathematical Equatings

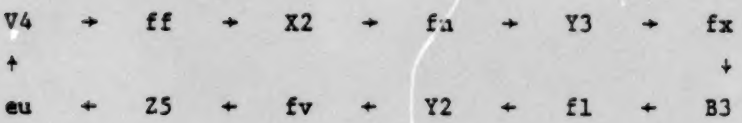


Figure 1: Verbal and Mathematical equating chains. Upper-case letters designate SAT forms and lower-case letters, anchor tests.

Table 4

Raw Score Summary Statistics for Verbal and Mathematical Equating Samples

Form[a]	Admin Date	N	SAT Form		Anchor Test		Covariance
			Mean	SD	Mean	SD	
Verbal Samples							
V4-fe	12/73	2665	35.04	16.37	14.01	8.54	123.22
X2-fe	4/75	2686	35.24	15.27	13.65	7.95	104.84
X2-fm	4/75	2562	34.42	15.31	16.74	8.07	106.75
Y3-fm	6/76	2578	34.48	16.34	16.14	8.41	121.17
Y3-fw	1/78	2549	31.37	15.86	14.36	8.17	114.28
B3-fw	5/79	2700	36.40	15.80	16.38	8.06	111.58
B3-fk	5/79	2665	35.90	15.24	15.04	8.01	106.07
Y2-fk	4/76	2879	34.16	14.84	15.08	8.19	105.23
Y2-fu	4/76	2774	33.57	14.50	16.02	7.44	92.29
Z5-fu	12/77	2853	30.73	15.61	14.43	7.69	103.94
Z5-et	12/77	2814	31.13	15.91	13.76	7.83	108.49
V4-et	12/73	2670	34.66	16.11	15.04	7.94	110.41
Mathematical Samples							
V4-ff	12/73	2628	20.74	12.89	9.46	5.55	59.91
X2-ff	4/75	2629	24.42	12.88	9.97	5.45	58.97
X2-fn	4/75	2527	23.84	12.90	9.73	5.73	63.24
Y3-fn	6/76	2553	24.05	13.30	9.57	5.85	67.40
Y3-fx	1/78	2455	21.48	13.74	8.47	6.33	77.11
B3-fx	5/79	2633	24.70	13.33	10.14	6.10	70.09
B3-fl	5/79	2596	24.30	12.97	9.46	5.36	59.09
Y2-fl	4/76	2815	23.25	13.49	9.13	5.46	62.83
Y2-fv	4/76	2721	23.68	13.46	9.76	6.01	69.98
Z5-fv	12/77	2774	20.52	13.49	8.38	5.96	69.73
Z5-eu	12/77	2739	20.13	13.38	8.53	5.92	68.77
V4-eu	12/73	2673	20.55	12.96	9.46	6.06	67.61

[a]Upper-case letters designate SAT forms and lower-case letters, anchor tests.

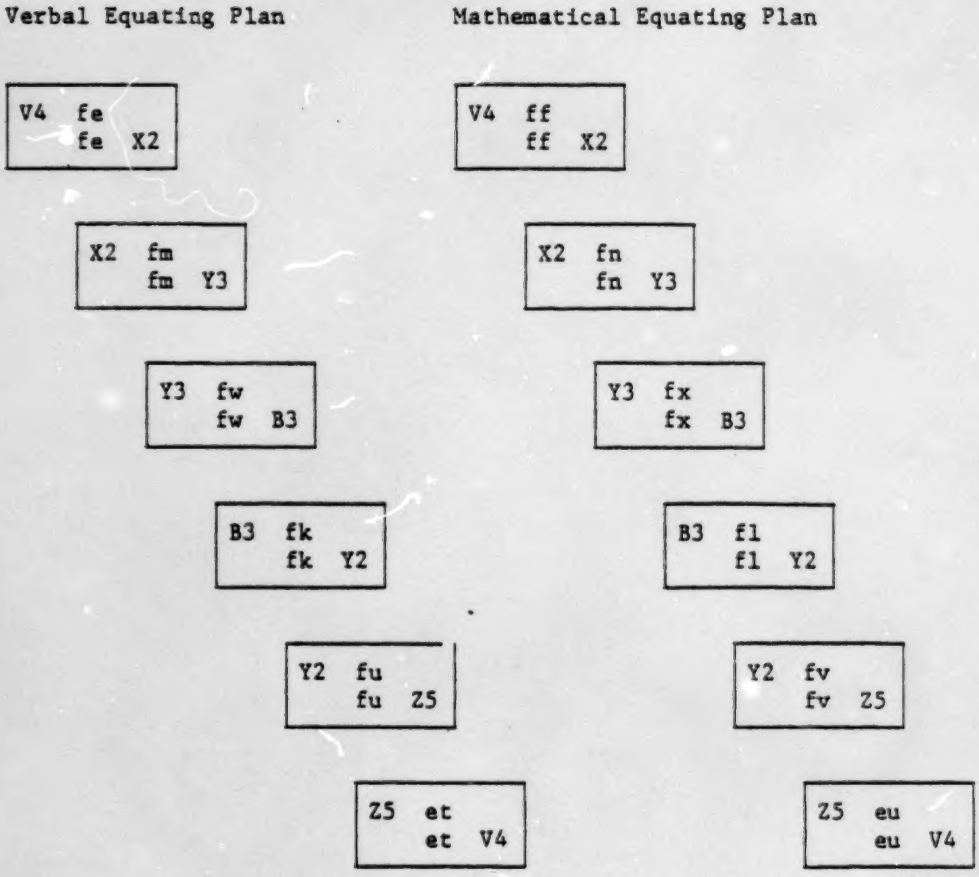


Figure 2: Verbal and Mathematical equating plans. Boxes indicate separate equating runs. Each pairing of an SAT form and an anchor test represents a sample of approximately 2670 cases.

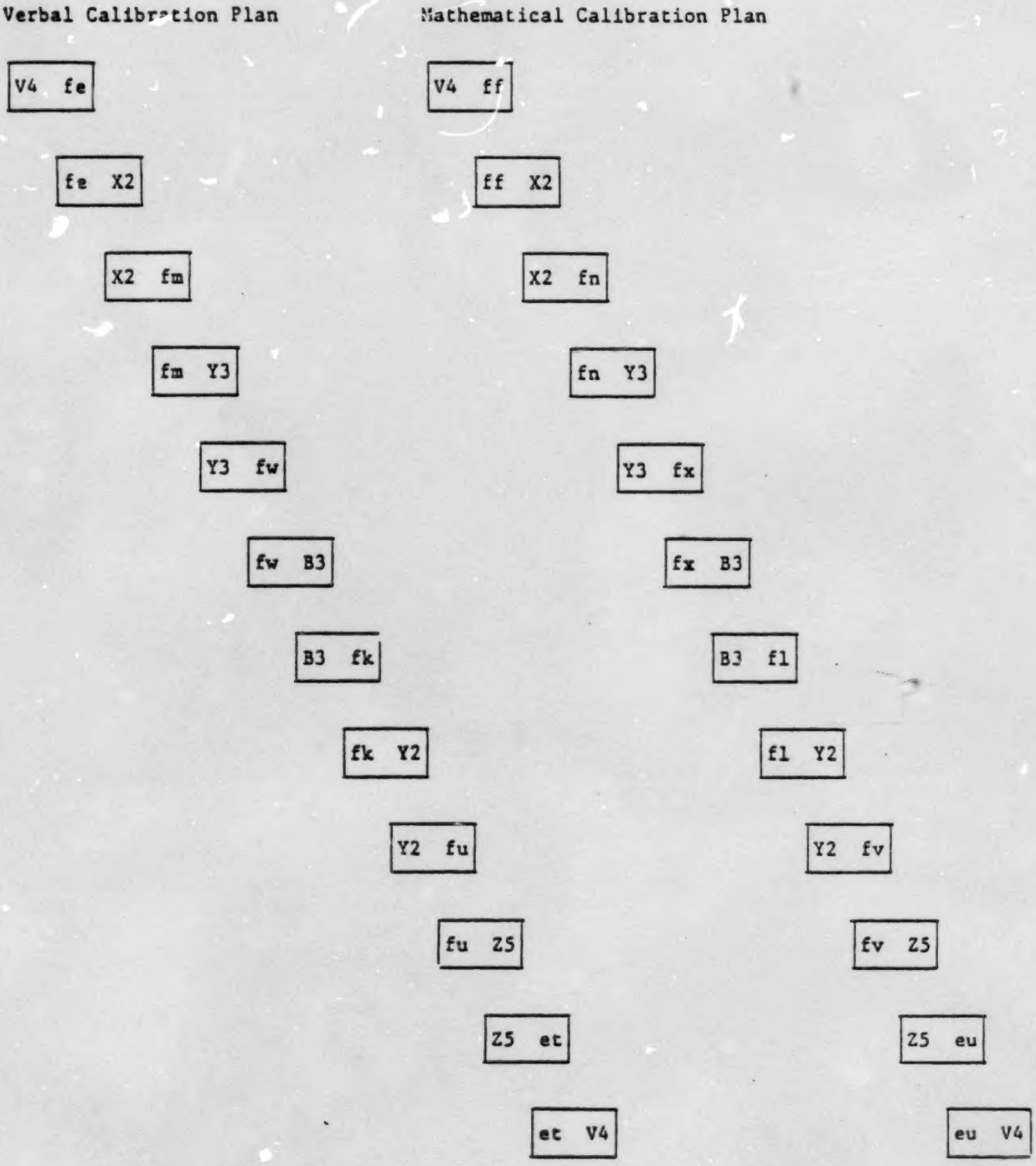


Figure 3: Verbal and Mathematical partial pre-calibration plans. Boxes indicate separate calibration runs. Each pairing of an SAT form and an anchor test represents a sample of approximately 2670 cases.

starting with the V4-fe box in the verbal chain and with the V4-ff box in the mathematical chain in Figure 3. For the second LOGIST run in the verbal chain, the item difficulties for the fe section obtained from the first run were used as input and were not re-estimated. The parameter estimates for X2 are then automatically placed on the same scale as those for fe. This procedure was continued down each chain with b's from items common with the previous LOGIST run being held fixed, placing the new items for the current run on the same scale. When each chain is complete, all parameter estimates within a chain are now on the same scale, and equating of scores on each pair of test editions can be performed.

In the other partial pre-calibration procedure (equated b's), each pairing of an SAT edition and an anchor test shown in Figure 3 was calibrated in an independent LOGIST run. To place all estimates of item parameters within a chain on a common scale, a sequential transformation process was employed. For example, the first two LOGIST runs for the verbal chain produced estimates for V4-fe on one scale, and for fe-X2 on a different scale. The two separate item difficulty estimates for items in section fe were then related by mean and sigma equating using weighted estimates of the mean and standard deviation. The weights are a combination of the standard error of the estimated item difficulties and weights designed to decrease the influence of outliers. This procedure yields the transformation that needs to be applied to the a and b parameters for all items from the fe-X2 calibration to place them on the same scale as V4-fe. The procedure is repeated, using the transformation obtained from the relationship between the difficulty estimates for edition X2 to place X2-fm items on the same scale. The transformations were continued sequentially down each chain resulting in a common scale for all item parameter estimates within a chain. Equating of each pair of test editions can then be performed.

The initial and final raw-score to scaled-score transformations for edition V4 are listed in Table 5 for SAT-Verbal and in Table 6 for SAT-Mathematical. Also given in Tables 5 and 6 are the raw score frequencies for the total group taking edition V4 in December 1973. The equipercentile conversions do not extend over the possible raw score range because equipercentile equating can only be done over that score range for which there are observations in the equating samples. Theoretically, the initial and final conversion of V4 raw scores to the College Board scale should be identical for each equating method.

Again, the weighted mean square difference was selected for use as a summary index to aid in the evaluation of the effectiveness of the various equating models. Examination of the SAT-Verbal data in Table 7 indicates that the total error was largest for the equipercentile method and smallest for the IRT equated b's method. All methods, except for the IRT equated b's method, overestimated the criterion mean. The best estimate of the mean was given by the IRT equated b's method and the worst estimate by the Tucker method. All methods overestimated the criterion standard deviation. The best estimate of the standard deviation was given by the IRT equated b's method and the worst estimate by the equipercentile method. All of the methods, except for the equipercentile method, produced similar estimates for the standard deviation. Bias accounted for at least 78 percent of the total error for all methods except the IRT equated b's method for which it accounted for only 21 percent of the total error. Results were similar for the three linear models. Of the linear methods, the Levine Equally Reliable model resulted in the smallest total error and the best estimates of the criterion mean and standard deviation. All of the IRT methods produced substantially better results than the linear and equipercentile methods. Results were similar for the IRT concurrent and fixed b's

TABLE 5

INITIAL AND FINAL TRANSFORMATIONS
OF SAT-VERBAL FORM V4 RAW SCORES TO SCALED SCORES

RAW SCORE	# PEOPLE	INITIAL SCALED SCORE	ESTIMATED SCALED SCORE						
			TUCKER	LEVINE EQ REL	LEVINE UNEQ REL	EQUITILE	CONCURRENT	FIXED B	EQUATED B
90	1	822.62	856.64	850.77	853.53		822.61	822.61	822.61
89	4	815.63	849.41	843.61	846.32		821.64	816.25	815.77
88	11	808.63	842.18	836.44	839.11		818.56	810.25	809.05
87	1	801.64	834.94	829.28	831.90		814.25	804.36	802.43
86	20	794.65	827.71	822.11	824.68		809.20	798.52	795.87
85	44	787.66	820.47	814.95	817.47		803.69	792.69	789.35
84	71	780.66	813.24	807.78	810.26		797.89	786.83	782.84
83	98	773.67	806.01	800.62	803.05		791.90	780.95	776.34
82	38	766.68	798.77	793.46	795.84		785.78	775.03	769.82
81	135	759.68	791.54	786.29	788.63		779.55	769.05	763.27
80	172	752.69	784.30	779.13	781.42	772.34	773.20	763.02	756.70
79	215	745.70	777.07	771.96	774.21	763.81	766.75	756.93	750.09
78	251	738.70	769.84	764.80	767.00	754.02	760.18	750.78	743.45
77	164	731.71	762.60	757.63	759.79	744.93	753.51	744.56	736.77
76	294	724.72	755.37	750.47	752.58	738.98	746.75	738.28	730.05
75	360	717.72	748.13	743.30	745.36	724.65	739.88	731.94	723.28
74	441	710.73	740.90	736.14	738.15	718.63	732.92	725.54	716.47
73	533	703.74	733.67	728.98	730.94	715.54	725.87	719.07	709.62
72	674	696.75	726.43	721.81	723.73	712.83	718.75	712.54	702.72
71	540	689.75	719.20	714.65	716.52	706.19	711.54	705.94	695.77
70	667	682.76	711.96	707.48	709.31	696.40	704.30	699.27	688.77
69	791	675.77	704.73	700.32	702.10	693.32	696.99	692.55	681.73
68	913	668.77	697.50	693.15	694.89	690.87	689.63	685.75	674.65
67	650	661.78	690.26	685.99	687.68	688.63	682.24	678.89	667.52
66	1002	654.79	683.03	678.83	680.47	686.40	674.81	671.97	660.34
65	1166	647.79	675.79	671.66	673.26	684.37	667.34	664.99	653.13
64	1313	640.80	668.56	664.50	666.04	681.92	659.87	657.94	645.88
63	1486	633.81	661.33	657.33	658.83	677.93	652.37	650.84	638.60
62	1022	626.81	654.09	650.17	651.62	665.21	644.86	643.68	631.29
61	1562	619.82	646.86	643.00	644.41	646.12	637.34	636.47	623.95
60	1801	612.83	639.62	635.84	637.20	636.46	629.81	629.21	616.59
59	1955	605.84	632.39	628.68	629.99	627.93	622.26	621.91	609.22
58	2162	598.84	625.16	621.51	622.78	622.69	614.73	614.56	601.83
57	1833	591.85	617.92	614.35	615.57	617.93	607.19	607.18	594.41
56	2228	584.86	610.69	607.18	608.36	611.22	599.68	599.78	587.02
55	2301	577.86	603.45	600.02	601.15	604.72	592.20	592.35	579.61
54	2094	570.87	596.22	592.85	593.94	597.30	584.70	584.89	572.19
53	2587	563.88	588.99	585.69	586.73	587.02	577.24	577.43	564.79
52	2202	556.88	581.75	578.53	579.51	578.07	569.80	569.95	557.38
51	2133	549.89	574.52	571.36	572.30	572.06	562.39	562.47	549.99
50	1674	542.90	567.28	564.20	565.09	566.39	555.01	555.00	542.62
49	1753	535.90	560.05	557.03	557.88	561.08	547.67	547.53	535.25
48	1998	528.91	552.82	549.87	550.67	555.48	540.36	540.07	527.91
47	1911	521.92	545.58	542.70	543.46	549.82	533.09	532.63	520.59
46	1894	514.92	538.35	535.54	536.25	545.06	525.89	525.22	513.29
45	1618	507.93	531.11	528.38	529.04	541.57	518.71	517.83	506.03
44	1739	500.94	523.88	521.21	521.83	538.07	511.59	510.46	498.78

TABLE 5 (CONTINUED)

INITIAL AND FINAL TRANSFORMATIONS
OF SAT-VERBAL FORM V4 RAW SCORES TO SCALED SCORES

RAW SCORE	A FREQ	INITIAL SCALED SCORE	ESTIMATED SCALED SCORE						
			TUCKER	LEVINE FQ REL	LEVINE UNEQ REL	EQUITILE	CONCURRENT	FIXED B	EQUATED B
43	4448	493.95	516.65	514.05	514.62	531.08	504.49	503.17	491.56
42	3793	480.95	509.41	506.88	507.41	523.60	497.44	495.82	484.37
41	5032	479.96	502.18	499.72	500.19	515.27	490.42	488.55	477.22
40	5449	472.97	494.94	492.55	492.98	504.30	483.44	481.32	470.09
39	5492	465.97	487.71	485.39	485.77	492.69	476.48	474.12	462.99
38	5435	458.98	480.48	478.23	478.56	479.96	469.56	466.96	455.91
37	5420	451.99	473.24	471.06	471.35	470.94	462.64	459.83	448.87
36	5794	444.99	466.01	463.90	464.14	463.25	455.74	452.73	441.84
35	5542	438.00	458.77	456.73	456.93	455.90	448.85	445.66	434.84
34	5933	431.01	451.54	449.57	449.72	445.48	441.95	438.62	427.86
33	6072	424.01	444.31	442.40	442.51	438.35	435.04	431.61	420.90
32	4537	417.02	437.07	435.24	435.30	431.38	428.11	424.61	413.95
31	5860	410.03	429.84	428.08	428.09	421.32	421.16	417.64	407.02
30	5996	403.04	422.60	420.91	420.87	414.64	414.14	410.68	400.10
29	6005	396.04	415.37	413.75	413.66	405.20	407.14	403.74	393.20
28	5990	389.05	408.14	406.58	406.45	396.60	400.05	396.81	386.30
27	4521	382.06	400.90	399.42	399.24	387.58	392.93	389.88	379.42
26	5359	375.06	393.67	392.25	392.03	379.61	385.76	382.96	372.54
25	5381	368.07	386.43	385.09	384.82	374.43	378.54	376.04	365.68
24	5325	361.08	379.20	377.93	377.61	370.10	371.25	369.11	358.81
23	5002	354.08	371.97	370.76	370.40	366.18	363.92	362.18	351.96
22	3715	347.09	364.73	363.60	363.19	361.08	356.55	355.26	345.12
21	4666	340.10	357.50	356.43	355.98	354.22	349.10	348.32	338.28
20	4491	333.10	350.26	349.27	348.77	346.81	341.60	341.38	331.44
19	4356	326.11	343.03	342.10	341.55	340.52	334.05	334.43	324.64
18	4086	319.12	335.80	334.94	334.34	334.99	326.44	327.46	317.83
17	2933	312.12	328.56	327.78	327.13	329.82	318.77	320.49	311.02
16	3472	305.13	321.33	320.61	319.92	323.10	311.05	313.49	304.22
15	3320	298.14	314.09	313.45	312.71	316.88	303.29	306.48	297.43
14	3112	291.15	306.86	306.28	305.50	308.42	295.48	299.45	290.64
13	2436	284.15	299.63	299.12	298.29	298.77	287.64	292.39	283.85
12	1967	277.16	292.39	291.95	291.08	290.73	279.80	285.30	277.06
11	2315	270.17	285.16	284.79	283.87	281.70	271.99	278.18	270.28
10	2209	263.17	277.92	277.63	276.66	271.42	264.18	271.02	263.50
9	1967	256.18	270.69	270.46	269.45	264.22	256.45	263.81	256.72
8	1754	249.19	263.46	263.30	262.23	258.84	248.78	256.55	249.94
7	1123	242.19	256.22	256.13	255.02	253.59	241.23	249.24	243.16
6	1240	235.20	248.99	248.97	247.81	233.52	233.81	241.88	236.39
5	1196	228.21	241.75	241.80	240.60	225.62	226.54	234.44	229.61
4	1057	221.21	234.52	234.64	233.39	220.72	219.44	226.93	222.82
3	853	214.22	227.29	227.48	226.18	202.96	212.52	219.33	216.03
2	606	207.23	220.05	220.31	218.97	193.38	205.78	211.64	209.23
1	522	200.24	212.82	213.15	211.76	191.70	199.22	203.85	202.41
0	444	193.24	205.58	205.93	204.55		192.86	195.94	195.56
-1	353	186.25	198.35	198.72	197.34		186.61	187.91	188.68
-2	230	179.26	191.12	191.65	190.13		180.59	179.75	181.78
-3	117	172.26	183.88	184.49	182.91		174.70	171.44	174.84

TABLE 5 (CONTINUED)

INITIAL AND FINAL TRANSFORMATIONS
OF SAT-VERBAL FORM V4 RAW SCORES TO SCALED SCORES

RAW SCORE	A (FREQ)	INITIAL SCALED SCORE	ESTIMATED SCALED SCORE						
			TUCKER	LEVINE EQ REL	LEVINE UNEQ REL	EQUITILE	CONCURRENT	FIXED A	EQUATED A
-4	129	185.27	176.65	177.93	175.70		168.97	162.97	167.87
-5	91	158.28	169.41	170.15	168.49		163.67	154.26	160.90
-6	52	151.28	162.18	163.00	161.28		156.46	146.75	153.89
-7	30	144.25	154.95	155.83	154.07		149.24	139.89	146.80
-8	6	137.30	147.71	148.67	146.86		142.01	133.03	139.71
-9	14	130.50	140.48	141.50	139.65		134.79	126.17	132.62
-10	6	123.31	133.24	134.34	132.44		127.56	119.32	125.54
-11	3	116.32	126.01	127.18	125.23		120.34	112.46	118.45
-12	2	109.32	118.78	120.01	118.02		113.11	105.60	111.36
-13	0	102.33	111.54	112.85	110.81		105.89	98.74	104.27
-14	1	95.34	104.31	105.68	103.59		98.66	91.88	97.18
-15		88.35	97.07	98.52	96.38		91.44	85.02	90.09
-16		81.35	89.84	91.35	89.17		84.21	78.16	83.00
-17		74.36	82.61	84.19	81.96		76.99	71.30	75.92
-18		67.37	75.37	77.03	74.75		69.76	64.44	68.83
-19		60.37	68.14	69.86	67.54		62.54	57.58	61.74
-20		53.38	60.90	62.70	60.33		55.31	50.73	54.65
-21		46.39	53.67	55.53	53.12		48.09	43.87	47.56
-22		39.39	46.44	48.37	45.91		40.86	37.01	40.47

A
SAT-VERBAL FORM V4 RAW SCORE FREQUENCY DISTRIBUTION FOR INITIAL DECEMBER 1973 ADMINISTRATION.

TABLE 6

INITIAL AND FINAL TRANSFORMATIONS
OF SAT-MATH FORM V4 RAW SCORES TO SCALED SCORES

RAW SCORE	A FREQ	INITIAL SCALED SCORE	ESTIMATED SCALED SCORE						
			TUCKER	LEVINE EQ REL	LEVINE UNEQ REL	EQUITILE	CONCURRENT	FIXED A	EQUATED A
60	14	805.63	817.43	812.74	815.32		805.62	805.62	805.62
59	114	796.74	808.47	803.78	806.30		799.33	798.79	796.40
58	205	787.86	799.50	794.82	797.28		792.03	791.76	786.76
57	47	778.97	790.54	785.87	788.26		784.29	784.64	777.15
56	317	770.09	781.57	776.91	779.24		776.37	777.43	767.66
55	475	761.21	772.61	767.96	770.22	759.96	768.37	770.11	758.29
54	638	752.32	763.64	759.00	761.20	756.23	760.32	762.67	749.00
53	795	743.44	754.68	750.05	752.18	753.30	752.23	755.09	739.78
52	930	734.56	745.72	741.09	743.16	747.35	744.03	747.35	730.59
51	878	725.67	736.75	732.14	734.14	731.45	735.73	739.44	721.43
50	1096	716.79	727.79	723.19	725.12	726.03	727.31	731.36	712.26
49	1312	707.90	718.82	714.22	716.10	717.05	718.77	723.12	703.09
48	1555	699.02	709.86	705.27	707.08	703.82	710.10	714.72	693.90
47	1144	690.14	700.89	696.31	698.06	693.33	701.32	706.15	684.70
46	1647	681.25	691.93	687.36	689.04	682.23	692.40	697.44	675.48
45	1939	672.37	682.96	678.40	680.02	669.17	683.37	688.59	666.23
44	2235	663.48	674.00	669.45	671.00	658.06	674.23	679.61	656.96
43	2452	654.60	665.03	660.49	661.98	648.83	664.98	670.52	647.67
42	2117	645.72	656.07	651.54	652.96	640.56	655.65	661.32	638.36
41	2588	636.83	647.10	642.58	643.94	633.13	646.23	652.02	629.01
40	2859	627.95	638.14	633.63	634.92	626.53	636.76	642.63	619.65
39	3213	619.06	629.18	624.67	625.90	619.06	627.24	633.17	610.26
38	3560	610.18	620.21	615.71	616.88	612.85	617.69	623.65	600.85
37	3059	601.30	611.25	606.76	607.86	606.54	608.13	614.08	591.43
36	3838	592.41	602.28	597.80	598.84	600.85	598.57	604.47	582.00
35	4022	583.53	593.32	588.85	589.82	595.79	589.02	594.84	572.57
34	4385	574.65	584.35	579.89	580.80	588.86	579.49	585.18	563.15
33	4761	565.76	575.39	570.94	571.78	577.93	570.01	575.53	553.74
32	3817	556.88	566.42	561.98	562.76	566.56	560.59	565.88	544.34
31	5353	547.99	557.46	553.03	553.74	553.41	551.23	556.25	534.97
30	5248	539.11	548.49	544.07	544.72	540.53	541.96	546.64	525.62
29	5511	530.23	539.53	535.11	535.70	529.07	532.77	537.07	516.30
28	5870	521.34	530.57	526.16	526.68	517.97	523.67	527.55	507.00
27	4779	512.46	521.60	517.20	517.66	508.28	514.67	518.06	497.73
26	6034	503.57	512.64	508.25	508.64	500.46	505.75	508.63	488.48
25	6466	494.69	503.67	499.29	499.62	494.60	496.92	499.24	479.25
24	6576	485.81	494.71	490.34	490.60	489.63	488.17	489.89	470.03
23	6894	476.92	485.74	481.38	481.58	484.30	479.47	480.58	460.82
22	5470	468.04	476.78	472.43	472.56	477.10	470.83	471.31	451.61
21	6934	459.15	467.81	463.47	463.54	469.46	462.22	462.08	442.40
20	6844	450.27	458.85	454.52	454.52	461.73	453.64	452.86	433.18
19	7186	441.39	449.88	445.56	445.50	454.16	445.06	443.67	423.96
18	7208	432.50	440.92	436.60	436.48	443.34	436.48	434.50	414.71
17	5444	423.62	431.95	427.65	427.46	429.57	427.88	425.33	405.45
16	6309	414.73	422.99	418.69	418.44	419.62	419.25	416.17	396.17
15	6986	405.85	414.03	409.74	409.47	411.54	410.58	407.01	386.87
14	7014	396.97	405.06	400.78	400.40	402.21	401.86	397.85	377.56

TABLE 6 (CONTINUED)

INITIAL AND FINAL TRANSFORMATIONS
OF SAT-MATH FORM V4 RAW SCORES TO SCALED SCORES

RAW SCORE	A FREQ	INITIAL SCALED SCORE	ESTIMATED SCALED SCORE						
			TUCKER	LEVINE EQ FREQ	LEVINE UNEQ FREQ	EQUIPERCENTILE	CONCURRENT	FIXED A	EQUATED A
13	6696	394.34	396.13	391.83	391.38	387.11	391.08	388.69	368.24
12	5006	379.20	387.13	382.87	382.36	372.00	384.24	379.52	358.90
11	6025	370.12	378.17	373.92	373.34	364.45	375.32	370.34	349.57
10	5949	361.43	369.20	364.96	364.32	357.26	366.31	361.16	340.25
9	5583	352.55	360.24	356.00	355.30	351.30	357.19	351.98	330.95
8	5411	343.66	351.27	347.05	346.28	344.46	347.94	342.80	321.69
7	3576	334.78	342.31	338.09	337.26	338.42	338.55	333.63	312.47
6	4577	325.90	333.34	329.14	328.24	331.85	328.97	324.47	303.33
5	4155	317.31	324.38	320.18	319.22	324.92	319.20	315.32	294.21
4	3849	308.13	315.41	311.23	310.20	318.08	309.19	306.20	285.32
3	3405	299.24	306.45	302.27	301.18	309.19	298.95	297.11	276.51
2	2065	290.36	297.49	293.32	292.16	301.38	288.48	288.07	267.84
1	2454	281.48	288.52	284.36	283.14	294.36	277.84	279.07	259.35
0	2049	272.59	279.56	275.41	274.12	286.18	267.10	270.14	251.06
-1	1421	263.71	270.59	266.45	265.10	279.88	256.39	261.28	242.97
-2	1152	254.82	261.63	257.49	256.08	276.50	245.88	252.50	235.10
-3	456	245.94	252.66	248.54	247.06	273.75	235.72	243.79	227.45
-4	530	237.06	243.70	239.58	238.04	272.06	225.97	235.15	220.01
-5	324	228.17	234.73	230.63	229.02	270.11	216.40	226.51	212.74
-6	186	219.29	225.77	221.67	220.00	258.73	204.91	217.67	205.57
-7	97	210.41	216.80	212.72	210.98	244.08	200.43	208.25	197.94
-8	17	201.52	207.84	203.76	201.96	225.42	192.07	199.48	189.64
-9	17	192.64	198.87	194.81	192.94		183.71	190.70	181.35
-10	4	183.75	189.91	185.85	183.92		175.34	181.93	173.06
-11		174.87	180.95	176.89	174.89		166.98	173.15	164.77
-12		165.99	171.98	167.94	165.87		158.61	164.37	156.48
-13		157.10	163.02	158.98	156.85		150.25	155.60	148.18
-14		148.22	154.05	150.03	147.83		141.89	146.82	139.89
-15		139.31	145.09	141.07	138.81		133.52	138.05	131.60
-16		130.45	136.12	132.12	129.79				
-17		121.57	127.16	123.16	120.77				

A
SAT-MATH FORM V4 RAW SCORE FREQUENCY DISTRIBUTION FOR INITIAL DECEMBER 1973 ADMINISTRATION.

Table 7

Summary and Discrepancy Indices[a] for Equating Methods

Used in the SAT Scale Drift Study

Index	Initial Scale (Criterion)	Linear Equating			Curvilinear Equating			
						IRT Equating		
		Tucker	Levine Eq. Rel.	Levine Uneq. Rel.	Equi%ile	Con-Current	Fixed b's	Equated b's
SAT-Verbal								
Scaled Score:								
Mean	435.37	456.05	454.04	454.22	454.14	445.73	444.80	434.19
Standard Deviation	109.09	112.85	111.76	112.49	115.89	112.89	111.39	110.14
Total Error[b]		441.90	355.62	366.73	451.67	125.15	96.67	6.67
Bias		20.68	18.67	18.85	18.77	10.36	9.43	-1.18
SD of Difference		3.76	2.67	3.40	9.96	4.23	2.79	2.30
Absolute Difference:								
Maximum/Raw Score		31.61/80	26.44/80	28.73/80	44.12/63	22.19/74	17.18/66	6.02/71
Minimum/Raw Score		12.58/1	12.91/1	11.52/1	.49/4	.27/9	3.61/1	.10/12, 51
SAT-Mathematical								
Scaled Score:								
Mean	468.05	476.79	472.44	472.57	472.57	472.05	472.65	452.04
Standard Deviation	113.25	114.28	114.16	114.99	112.14	115.52	118.51	118.06
Total Error[b]		77.44	20.09	23.46	65.09	26.90	50.82	280.85
Bias		8.74	4.39	4.52	4.52	4.00	4.60	-16.02
SD of Differences		1.03	.91	1.73	6.69	3.31	5.45	4.94
Absolute Difference:								
Maximum/Raw Score		11.40/55	6.75/55	9.01/55	41.94/-5	14.48/-6	16.22/45	22.81/4
Minimum/Raw Score		6.32/-8	2.24/-8	.44/-8	.00/39	.29/3	.02/11	2.93/55

[a]Computed for SAT-Verbal raw scores 1 through 80 and for SAT-Mathematical raw scores -8 through 55.

[b]$(\text{SD of Difference})^2 + (\text{Bias})^2$. Bias equals $(M_i - M_c)$, where M_c refers to the mean criterion score and M_i refers to the mean yielded by a particular equating method.

methods. Of the three IRT methods, the equated b's method resulted in the smallest total error and the best estimates of the criterion mean and standard deviation.

Examination of the SAT-Mathematical data in Table 7 indicates that the total error was smallest for the Levine Equally Reliable model and largest for the IRT equated b's method. All methods, except for the IRT equated b's method, overestimated the criterion mean. The best estimate of the mean was given by the IRT concurrent method and the worst estimate (a substantial underestimate) by the IRT equated b's method. All methods, except for the equipercentile method, overestimated the criterion standard deviation. The best estimate of the standard deviation was given by the Levine Equally Reliable model and the worst by the IRT fixed b's method; however, the equipercentile model resulted in the largest standard deviation of the difference between the estimated and criterion scores. Of the three linear models, the Levine Equally Reliable model resulted in the smallest total error and the best estimates of the criterion mean and standard deviation; however, results were very similar for both Levine models. Of the three IRT methods, the IRT concurrent method resulted in the smallest total error and the best estimates of the criterion mean and standard deviation. The best IRT method (concurrent) when compared to the best conventional equating method (Levine Equally Reliable) produced a better estimate of the criterion mean but a worse estimate of the criterion standard deviation. Overall, results for the two Levine models and the IRT concurrent method were similar and did not differ greatly from the criterion.

To summarize, the results for verbal scores were somewhat different from those for mathematical scores. For equating SAT-Verbal scores, the three IRT methods were clearly more accurate than the three linear methods or the

equipercentile method. For equating SAT-Mathematical scores, two of the linear methods and one of the IRT methods were more accurate than the other methods. The largest inaccuracies occurred in the application of the linear and equipercentile methods to the verbal scores. For this reason, the authors recommended that IRT methods be used at least for SAT-Verbal scores.

Change in SAT Specifications

Due to a shortage of high-difficulty verbal questions, it became necessary to change the difficulty specifications for SAT-Verbal editions that were being constructed for use beginning in January 1982. Table 8 shows the previous specifications - those in effect from 1974 through 1981 - and the current specifications. Notice that the decrease in the number of very difficult questions is balanced by an increase in the number of moderately difficult questions and a decrease in the number of easy questions, so that the average difficulty of the test remains essentially the same.

Highly difficult questions for the SAT-Verbal have always been difficult to develop. Few such questions tend to survive the pretesting and revision process intact. In the past, the SAT-Verbal included some highly difficult questions that tested knowledge of specialized or infrequently used words. In recent years, such questions have not been included in the SAT. Test disclosure further compounded the problem. Until recently, the developers of the SAT-Verbal were able to re-use high-difficulty questions taken from previous editions of the SAT. Now the situation has changed. The College Board's decision to permit test-takers to obtain their test booklets and answer sheets from some administrations prevents the developers of the SAT from re-using test questions.

A study by Donlon and Marco, (1981) was designed to determine which type of equating - linear or IRT - would more accurately preserve the meaning of scaled scores on the SAT when the difficulty specifications were changed. The study

Table 8 Difficulty specifications for SAT-verbal: number of questions at each difficulty level.

Equated Delta	Specified Number of Questions	
	1974-81	1982-
17.0 or higher	2	0
16.0-16.9	4	2
15.0-15.9	10	6
14.0-14.9	10	14
13.0-13.9	6	10
12.0-12.9	6	8
11.0-11.9	6	7
10.0-10.9	8	7
9.0- 9.9	8	10
8.0- 8.9	10	8
7.0- 7.9	8	6
6.0- 6.9	4	4
5.9 or lower	3	3
Specified mean equated delta	11.4	11.4
Specified standard deviation of equated delta values	3.3	3.0

involved the comparison of SAT-Verbal score distributions computed in different ways for the same group of test-takers. The test-takers had taken an edition of the SAT that was built to the 1974 specifications. The test-takers' responses were used to estimate their raw scores on a simulated test built to the 1982 (interim) SAT-Verbal specifications. The raw scores on the simulated test were then converted to scaled scores by equating them to the actual test scores, thus simulating the effects of a change from the old to the new specifications. The equating was done by two methods - one linear method and one three-parameter IRT method - producing two distributions of scaled scores. The experiment was actually done four times: twice for each of two different editions of the SAT, once using the total group (general sample) of test-takers and once using a subsample (select sample) who had applied to selective colleges. An ideal equating procedure would have resulted in each student's receiving the same scaled score on the new (simulated) test as on the old (actual) test. The distributions of the test-takers' scaled scores on the new and old tests would have been identical. The results of the study are illustrated in Figures 4-7. The results showed that, in comparison with linear equating, IRT equating produced a scaled score distribution that more accurately reproduced the distribution of scaled scores on the older test, particularly so for the select samples. In general, linear equating tended to produce too few high scores (above 600). The authors concluded that IRT equating should be used to equate scores on SAT editions constructed to different difficulty specifications.

Feasibility of IRT Equating

A study was carried out by Petersen and Cook (1981) to investigate the feasibility of using concurrent IRT equating procedures operationally to equate SAT scores. More specifically, the study sought to refine the procedures for

FIGURE 4

PERCENTAGE OF SCORES FALLING AT OR BELOW DESIGNATED SCALED SCORE LEVELS IN THREE DISTRIBUTIONS: SIMULATED IRT, SIMULATED LINEAR, AND ACTUAL TEST (INTERIM SPECIFICATIONS) (GENERAL SAMPLE - DECEMBER 1977 FORM)

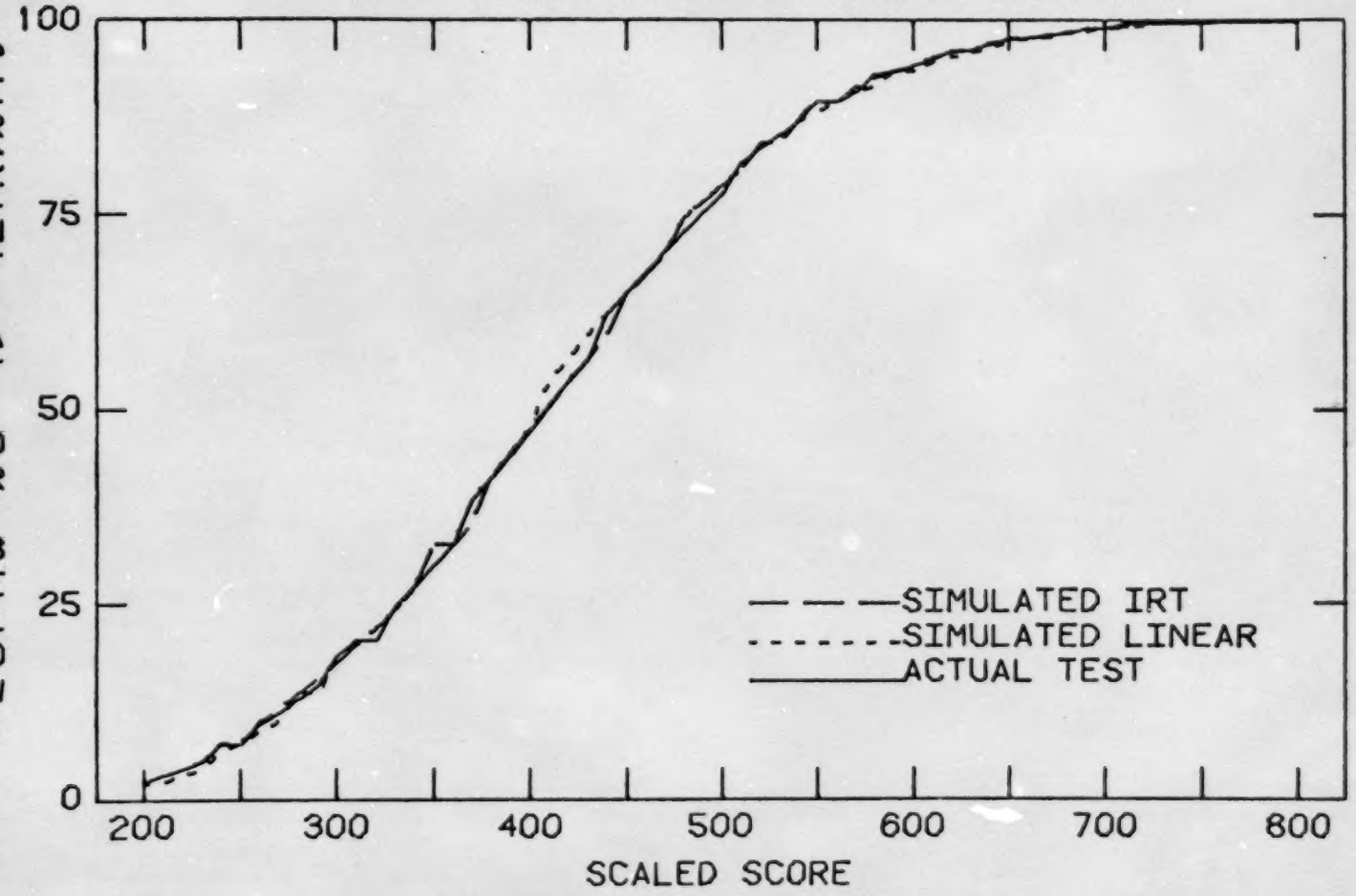


FIGURE 5

PERCENTAGE OF SCORES FALLING AT OR BELOW DESIGNATED SCALED SCORE LEVELS IN THREE DISTRIBUTIONS: SIMULATED IRT, SIMULATED LINEAR, AND ACTUAL TEST (INTERIM SPECIFICATIONS) (GENERAL SAMPLE - MAY 1979 FORM)

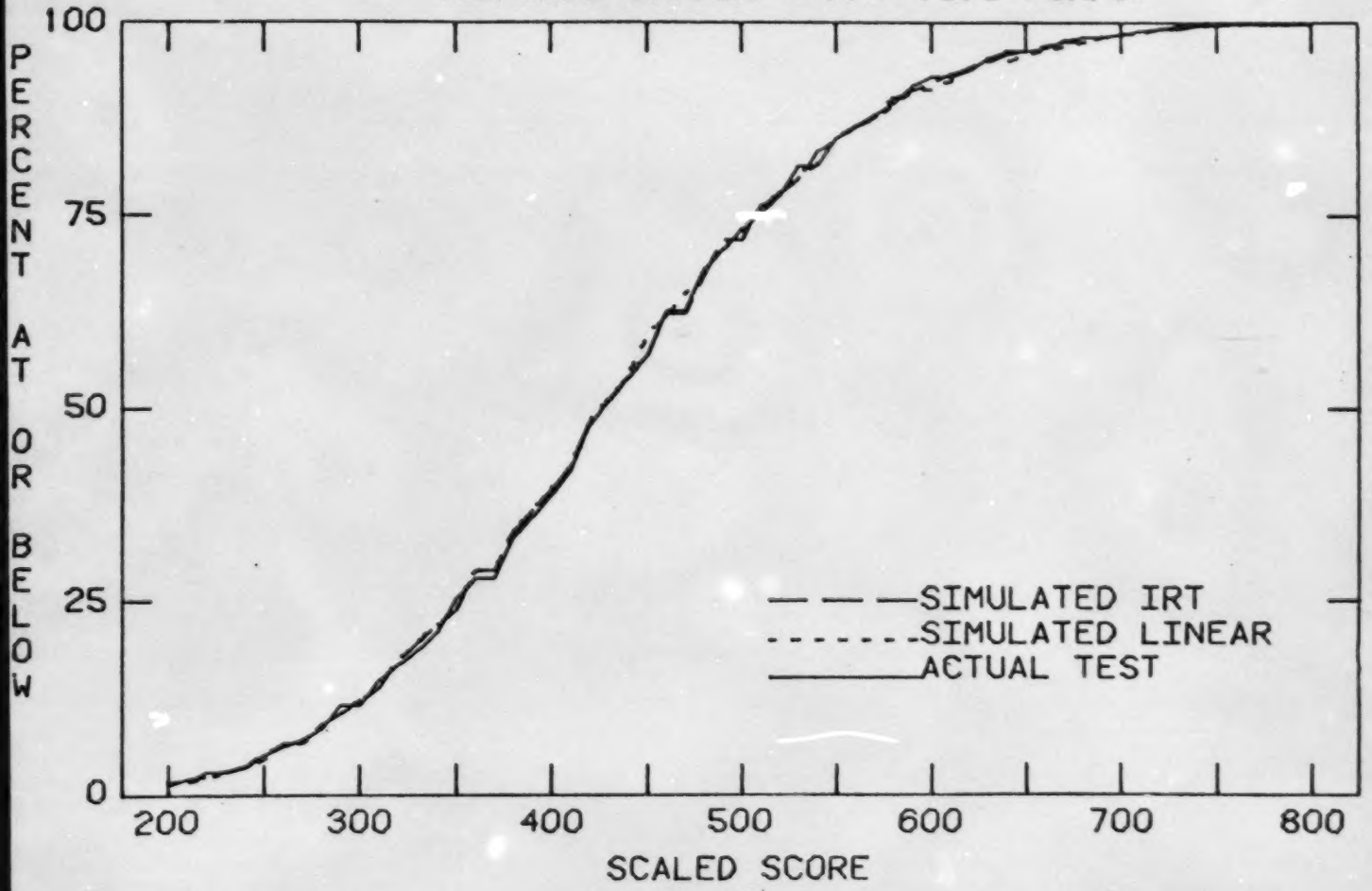


FIGURE 6

PERCENTAGE OF SCORES FALLING AT OR BELOW DESIGNATED SCALED SCORE LEVELS IN THREE DISTRIBUTIONS: SIMULATED IRT, SIMULATED LINEAR, AND ACTUAL TEST (INTERIM SPECIFICATIONS) (SELECT SAMPLE - DECEMBER 1977 FORM)

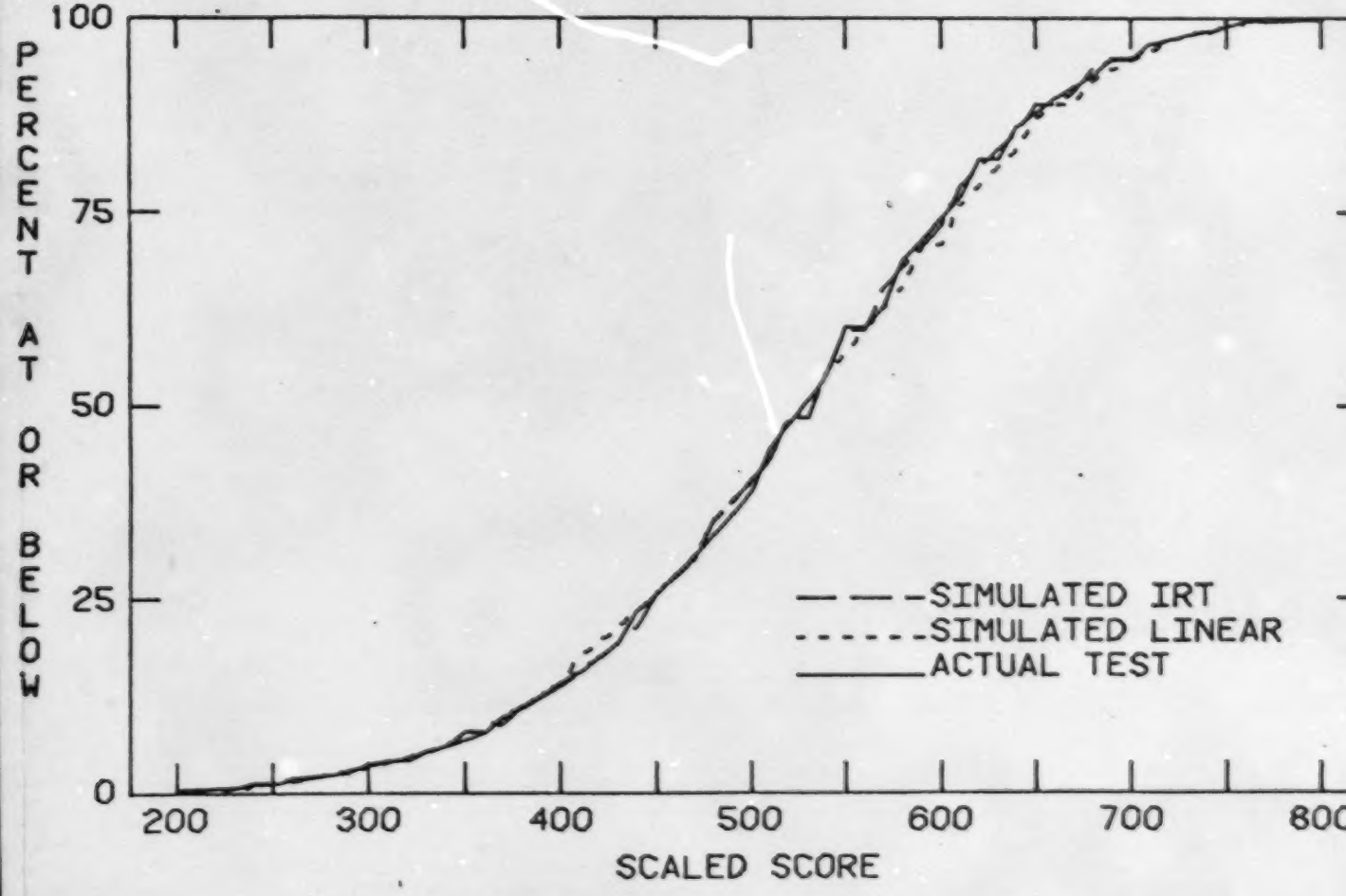
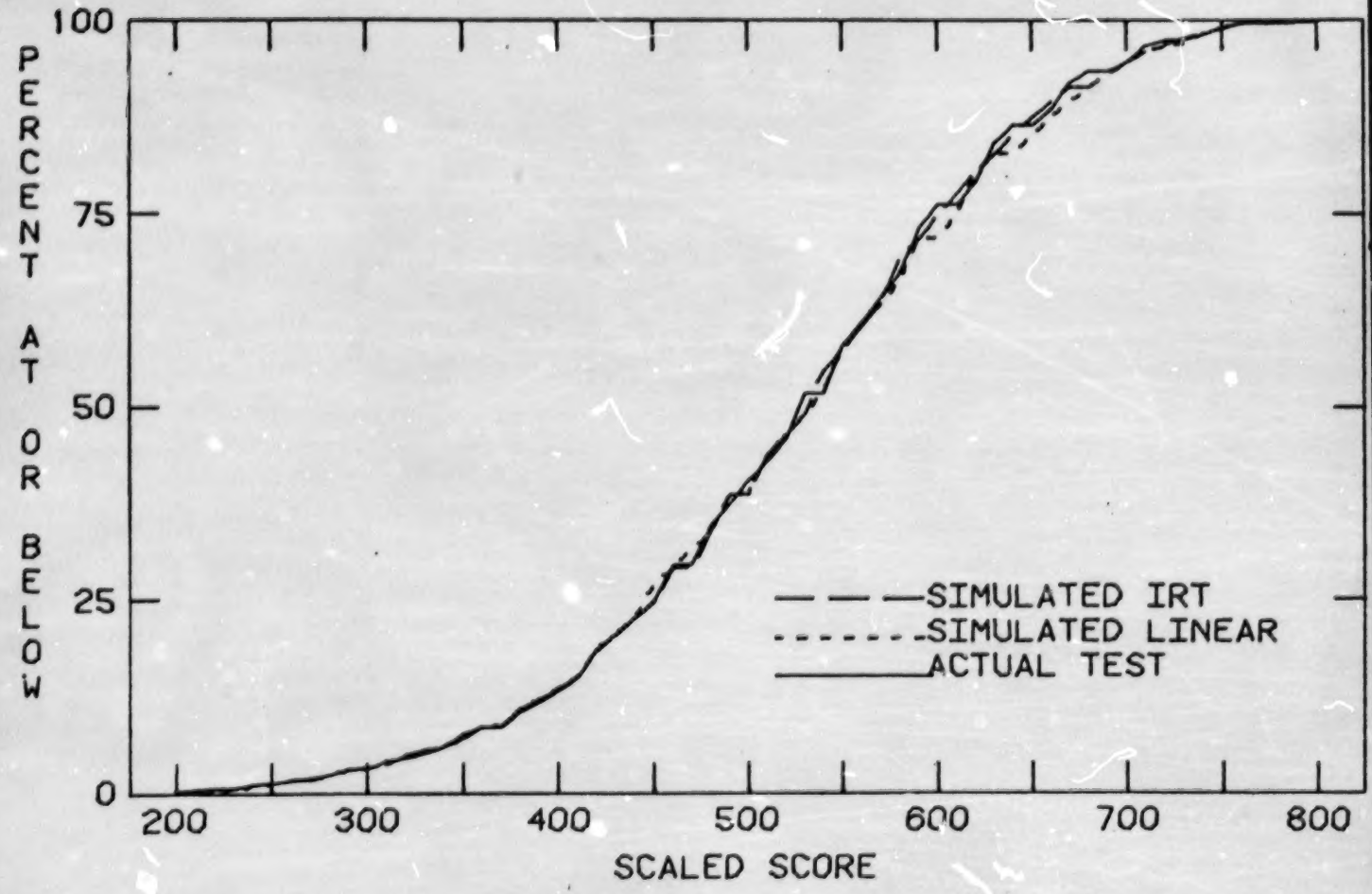


FIGURE 7

PERCENTAGE OF SCORES FALLING AT OR BELOW DESIGNATED SCALED SCORE LEVELS IN THREE DISTRIBUTIONS: SIMULATED IRT, SIMULATED LINEAR, AND ACTUAL TEST (INTERIM SPECIFICATIONS) (SELECT SAMPLE - MAY 1979 FORM)



operational use of concurrent IRT equating, to identify possible problems in using these procedures, to determine the cost associated with and schedule changes required for implementation of concurrent IRT equating, and to identify further developmental work required for the implementation.

The same data collection design is used for concurrent IRT equating of SAT scores as was used for linear equating, but the score relationships between the new and old editions of the SAT are established on the basis of item response data rather than score data. Each booklet containing an SAT test edition consists of six sections: two verbal sections (SAT-Verbal), two mathematical sections (SAT-Mathematical), a Test of Standard Written English, and an experimental section containing either an equating (anchor) test or a pretest. Each new edition is printed with 10 to 40 different experimental sections, which are spiraled when the edition is administered. For both SAT-Verbal and SAT-Mathematical, scores on two equating tests are used to link scores on the new edition to scores on two earlier editions, each of which was administered with one of the same equating tests. Four samples of data are used in the equating, one sample taking each combination of total test and equating test (see Figure 8). Item parameters for the three total tests and two equating tests are estimated and put on scale together in a single calibration run using the computer program LOGIST. Scores on the new edition are then equated to scores on each of the earlier editions and the results averaged.

The study included four trial runs of the concurrent IRT equating procedures, two for the SAT-Verbal only and two for both the SAT-Verbal and SAT-Mathematical. A large cost associated with IRT equating is the estimation of item and ability parameters. Thus, the IRT equating was done twice for one of the verbal tests using two different sample sizes: once using the operational samples used for

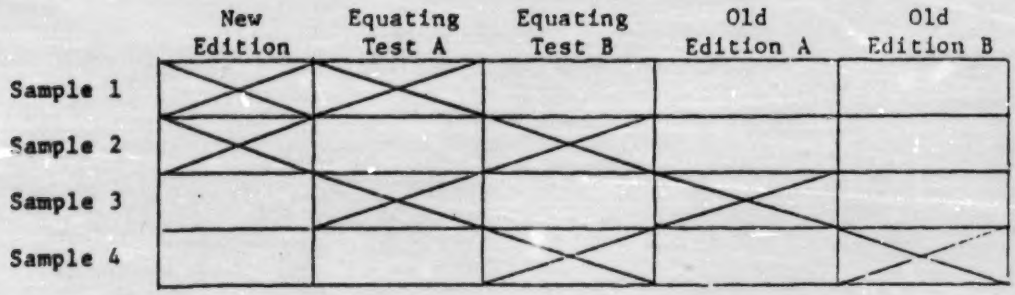


Figure 8. Operational calibration design for SAT-V or SAT-M (The "Xs" indicate the tests taken by the respective samples.)

linear equating (at least 5000 cases) and once using every other case from the operational samples. The two conversions and the resulting summary statistics were essentially the same, indicating that the use of samples of only 2500 cases would have a negligible effect on the stability of the IRT conversions. It was found that, in many instances, there were no discrepancies in SAT scaled-score conversions obtained from linear and IRT equating methods. The majority of the discrepancies that occurred were 10 points corresponding to a raw score difference of 1 point. Thus, it was anticipated that score users would not be likely to notice a change in score distributions due to a change in equating procedures. Overall, the study showed that IRT equating was feasible, that it would require a slight extension of the schedule for equating SAT scores, and that its increased cost would be moderate.

A follow-up study, the last in the series (Petersen and Whittington, 1981), investigated the effects of the change to IRT equating on the SAT scores of applicants to colleges of varying degrees of selectivity. The study compared the linear methods previously used to equate SAT scores with the IRT method now being used. The comparison focused on the differences in the score distributions produced by the two equating methods. The comparison was made for both verbal and mathematical scores in the full group of test-takers and in the group of applicants to each of 30 colleges. These colleges varied widely in the average academic ability of their applicants. They included the ten colleges whose applicants had the highest mean SAT scores (verbal plus mathematical), five colleges near the 75th percentile, five near the median, five near the 25th percentile, and the five lowest.

The differences between the score distributions produced by the two equating methods were small - generally smaller than the year-to-year differences in the

distribution of applicants' SAT scores. The authors concluded that, for the institutions studied, the impact of IRT would not have been discernible. In addition, the small differences between the two equating methods showed no consistent pattern, leading the authors to conclude that any slight increase in mean or standard deviation for one edition of the SAT would be likely to be cancelled out by a slight decrease for another edition. The results of this research indicate that college admissions officers are not likely to notice a change in the SAT scores of their applicants as a result of the change to IRT equating.

Studies in Progress

Research on IRT equating of the SAT that is currently being carried out in the College Board Division of Educational Testing Service is concerned principally with partial or full pre-calibration of SAT editions and/or anchor tests.

Because IRT item parameters are group independent, it is possible to calibrate (estimate item parameters) for a particular group of examinees (e.g., a pre test sample) and use these parameter estimates to obtain raw-score to scaled-score conversions for any group of examinees that might take a particular edition of the test. Thus, it is possible to estimate item parameters prior to the administration of the items in a new edition of the test (pre-calibration) and determine the raw-score to scaled-score conversions for the new edition prior to administration (pre-equating).

The description given above refers to full pre-calibration, i.e., all items in a particular test edition have been calibrated prior to administration. The situation can exist where some items have been pre-calibrated and some are calibrated when the test edition is administered the first time. This situation is referred to as partial pre-calibration.

The first study described in the remainder of this section is concerned with the full pre-calibration and subsequent pre-equating of verbal sections of the SAT. A similar study using SAT-Mathematical data will be carried out in the fall of 1982.

The second study focuses on the partial pre-calibration of the verbal and mathematical sections of the SAT. Work on this study began in August of 1982 and is scheduled for completion in June 1984.

IRT Equating with a Full Pre-calibration Design

As mentioned previously, pre-equating refers to the process of establishing conversions from raw to scaled scores prior to the time the new test is administered operationally. The process depends on the adequate pretesting of a pool of items from which the new test will be built, the calibration of these items using IRT methods prior to administration in a final form of the test (pre-calibration) and the utilization of a linking scheme to place the IRT parameters from the pretested items on the same scale. In light of recent test disclosure legislation, pre-equating offers a number of advantages: 1) IRT-based pre-equating is unaffected by the possible future problem of disclosing equating sections, because there would be no need for equating sections in the first place; 2) since equating using IRT pre-calibrated parameters is possible prior to the actual administration of the test, new editions can be introduced at low volume special administrations, a particular problem if traditional methods are used; and 3) pre-equating would actually permit a reduction in the usual score reporting cycle while allowing, at the same time, for more hours to do the equating itself.

The purpose of this study (Eignor, 1982) was to determine the extent to which item parameters estimated on SAT-Verbal pretest data can be used

for equating purposes in a situation where operational SAT testing data has normally been used. The items that appear in any final SAT-Verbal edition come from multiple pretests, and to the extent that the item statistics are sensitive to the context in which the item appears, or to sample differences, there may be differences between these statistics and statistics generated using data from the actual operational administration, or lack of fit between equating based on pretest item statistics and operational item statistics. More specifically, in this study, items from two operational SAT-Verbal editions were calibrated almost completely from pretest data. (Upon inspection, it was found that pretest data did not exist for eight of the eighty-five items in one verbal edition, and hence data from the operational administration of the test had to be used in the calibration system for these eight items.) An elaborate linkage system of other SAT-Verbal editions and equating sections, quite representative of the system that would exist were pre-equating to be considered for regular use, was devised and utilized to get parameter estimates for these items, contained in multiple pretests and the one intact edition, on the same scale. The two verbal editions under consideration were both part of this linkage system.

The effects of using these parameter estimates, estimated from the pretest data, on the equating process was evaluated in the following way. Each of the SAT-Verbal editions under study, when administered for the first time operationally, was equated to two different earlier editions. Conventional linear equating methods were used when this equating was done. The equating was re-done using item parameter estimates based on the pretest data and item parameter estimates generated from the operational administration. In each case, IRT true-score equating was performed. For each test edition, the IRT

equating based on pretest statistics was then compared to the IRT equating based on operational test data and the linear equating actually used when each edition was put on scale.

Pretest items for the verbal sections of two editions of the SAT, 3ASA3 and 3BSA3, were calibrated and placed on a common scale through a linkage system involving operational test editions administered in common with the pretests and equating sections in common with multiple operational editions. The calibration linkage system, involving the pretests, operational editions, and equating sections is depicted in Figure 9. Only those items in the pretests depicted that are contained in final editions 3ASA3 and 3BSA3 were calibrated, or in other words, only 170 verbal pretest items (85 items per edition times two editions) were calibrated. Responses from samples of approximately 3000 examinees taking each pretest-final test edition combination and approximately 2700 taking each final edition-equating section combination were used for calibration purposes.

Thirteen separate LOGIST runs were necessary to calibrate the pretest items, operational and equating section items used for linking purposes, and the editions to be used for equating purposes. These LOGIST runs are numbered 1-13 in Figure 9. Each of the separate LOGIST runs generated item parameter estimates on the particular scale defined by that run, and hence, a scaling program had to be run to put item parameter estimates from the separate LOGIST runs on a common scale.

The scalings are indicated by the arrows in Figure 9. A recently devised scaling method (Stocking and Lord, 1982) was used in the study. A linear transformation is found which places new edition item parameters on the scale of the old. The slope and intercept of the linear transformation

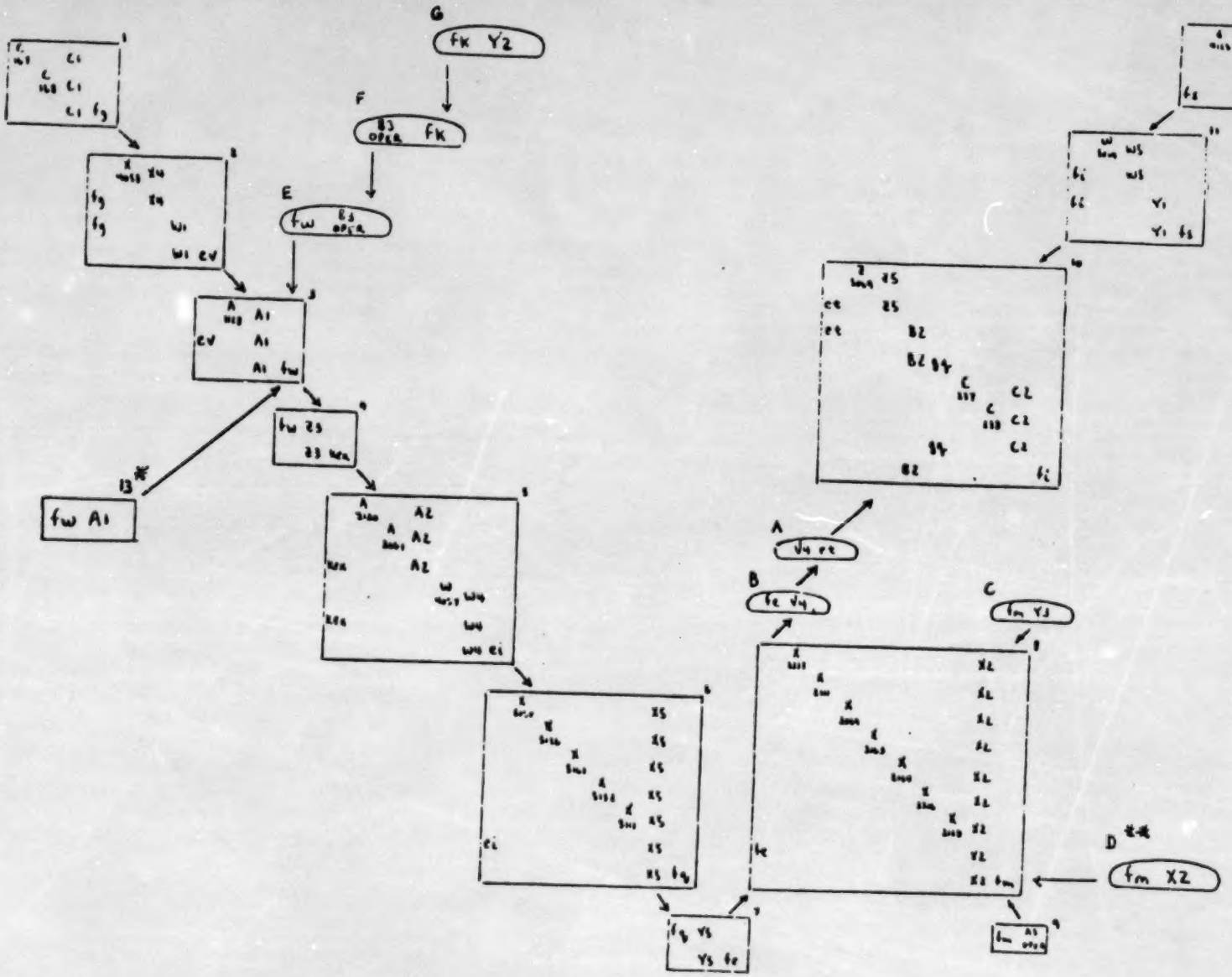


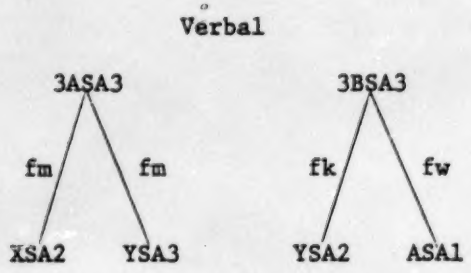
Figure 9: Verbal pre-equating and operational calibration and linking plan. Upper-case letters followed by one digit designate operational SAT forms. Upper-case letters followed by three or four digits designate pretest forms. Lower-case letters designate common item equating sections. Boxes (numbered 1-13) indicate separate LOGIST calibration runs. Ovals (lettered A-G) indicate forms and equating sections for which parameter estimates already exist from SAT IRT drift study. Arrows indicate direction in which scaling (linking) runs will take place.

* LOGIST run 13 necessary so that all 85 A1 items can be calibrated.

** Linking of D necessary so that the 45 X2 items not used in LOGIST run 8 can be placed on scale.

are chosen to minimize the average squared difference between number right true scores on the common item set for an arbitrary group of examinees who have taken the old edition of the test. This method implicitly makes use of information from all the parameters characterizing an item because number right true scores are used in the minimization process.

Operationally, the verbal sections of 3ASA3 and 3BSA3 were each equated to two old editions of the SAT, and this offers an ideal mechanism for evaluating the effects on equating of using items calibrated from pretest data. The following diagram depicts the actual equating that took place, and the common item sections used for the equatings.



For each equating depicted, IRT true-score equating was done three different ways. The first way involved the use of item parameter estimates based on pretest items which constitute 3ASA3 and 3BSA3, while the other two ways involve the use of item parameter estimates based on data collected when 3ASA3 and 3BSA3 were operationally administered in an intact fashion. The second and third way differ in the following manner. In one situation, referred to as operational equating, 3ASA3, 3BSA3, and the old editions to which they were equated were placed on the same scale by being linked into the overall calibration and linking plan shown in Figure 9. In this

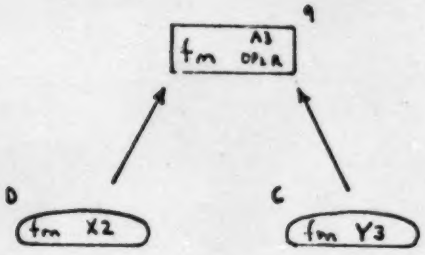
situation, the editions to be equated are linked indirectly through intervening LOGIST runs which contain multiple test editions and equating sections. In the other case, referred to as direct operational equating, the new (3ASA3 and 3BSA3) and old editions to be equated were linked directly through common equating sections. This linking is depicted in Figure 10 and represents possibly the best criterion against which to evaluate the pre-equating in that the relationship between parameter estimates, for the editions to be equated, from the separate LOGIST runs have not been influenced by intervening scalings.

In addition to the comparisons made involving the three varieties of IRT true-score equating, the effect of using pretest-based parameter estimates in the equating process was also evaluated by comparing the equatings to the linear equatings actually used when 3ASA3 and 3BSA3 were equated operationally.

Because each of the SAT-Verbal new editions was equated to two old editions, there are results for the single equatings (one for each new edition-old edition pairing) and subsequent results obtained by combining the two old edition equatings for each of the new editions. Only the results of the combined equatings will be discussed here. These results are presented in Figures 11-14. The figures contain plots of the formula-score to scaled-score conversions resulting from the various equatings as well as plots of residuals.

These residuals are simple differences between scaled scores resulting from the pre-equating and the IRT operational and direct operational equating, calculated for each possible formula-score point. The plots use the IRT operational (or direct operational) equating as the baseline and show differences between the baseline equating and the pre-equating results across the formula-score scale.

To equate A3 to X2 and Y3
directly using operational A3 data



To equate B3 to Y2 and A1
directly using operational B3 data

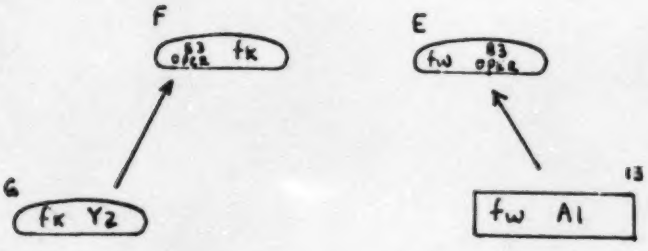


Figure 10: Verbal direct operational calibration and linking plan. Upper-case letters followed by one digit designate operational SAT forms. Lower-case letters designate common item equating sections. Boxes and ovals are numbered to directly correspond to comparable boxes and ovals in Figure 1. Arrows indicate direction in which scaling (linking) will take place.

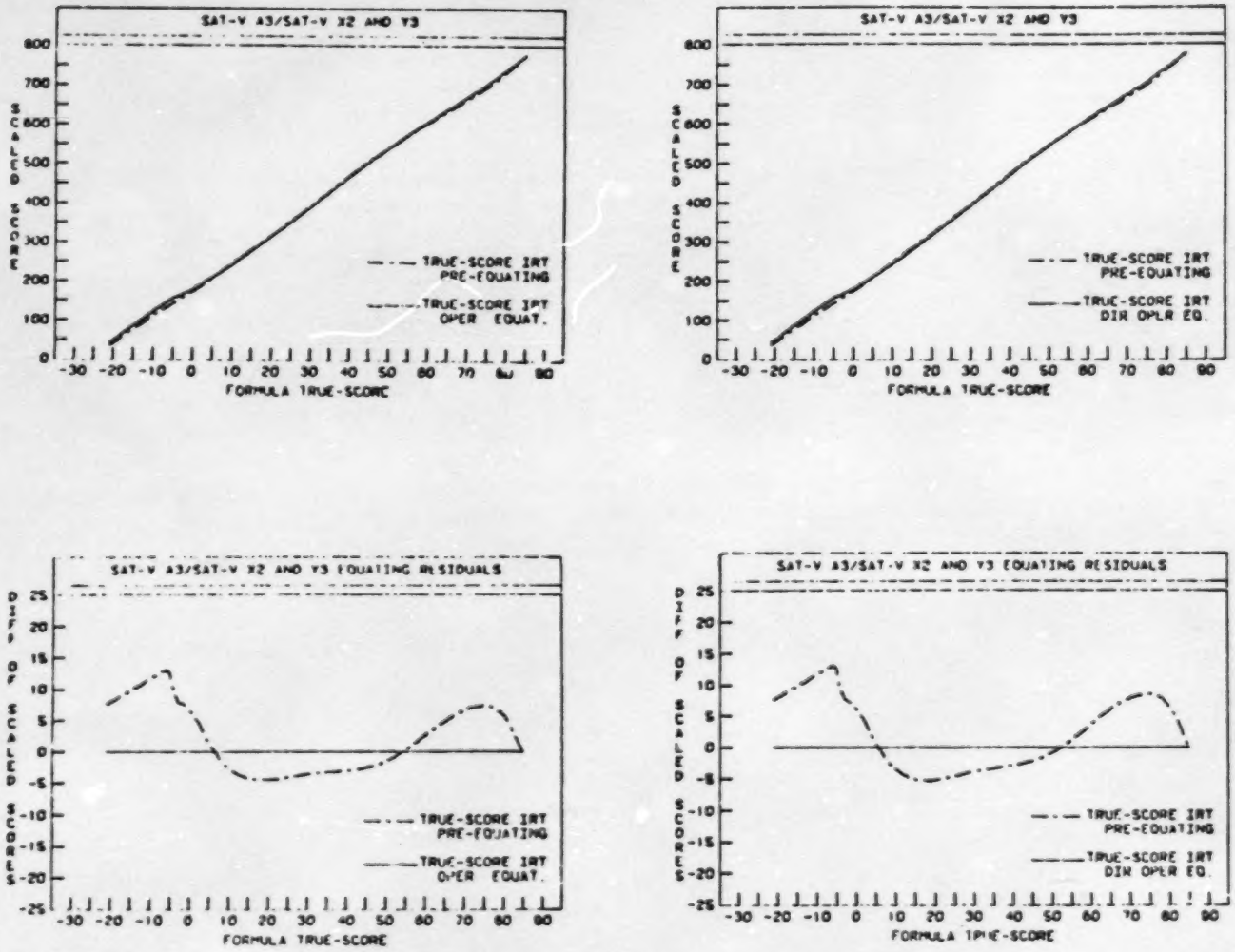
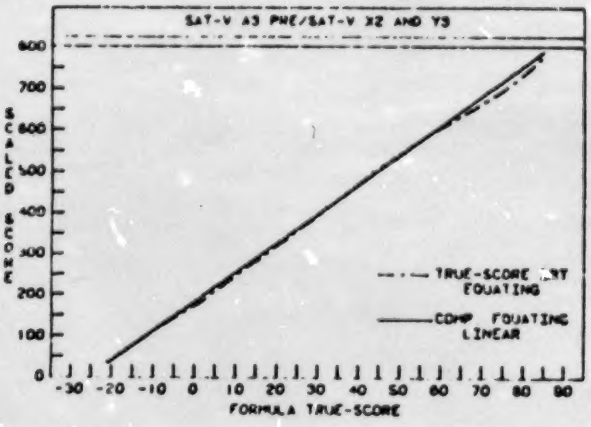
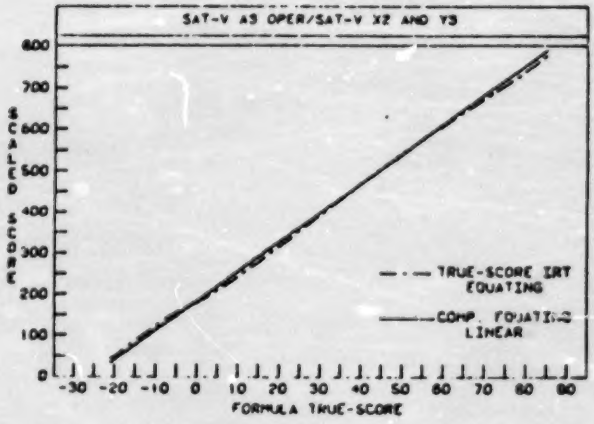


Figure 11: SAT-V Form 3ASA3 equated to SAT-V Form XSA2 and SAT-V Form YSA3 - Plots of 1) IRT pre-equating raw to scaled score transformation compared to corresponding IRT operational and direct operational equating raw score to scaled score transformations, and 2) Differences between scaled scores (IRT operational - IRT pre-equating) resulting from the equatings.





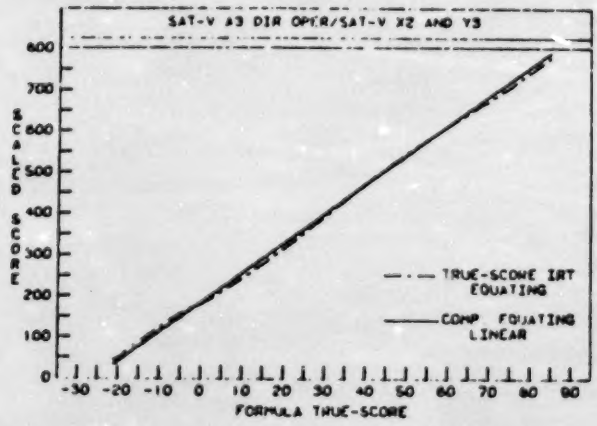
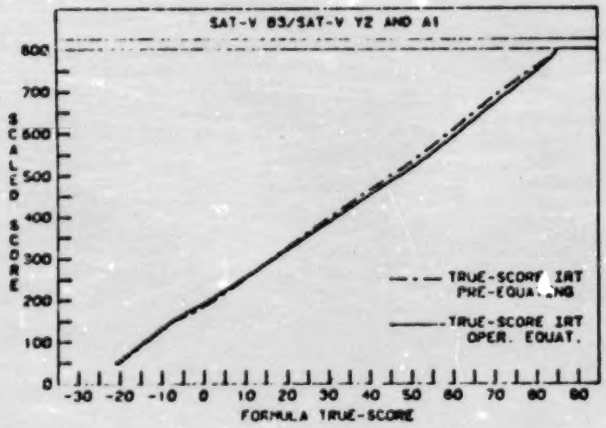
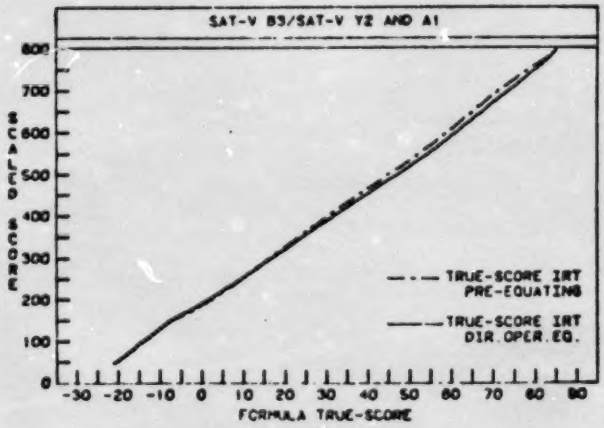
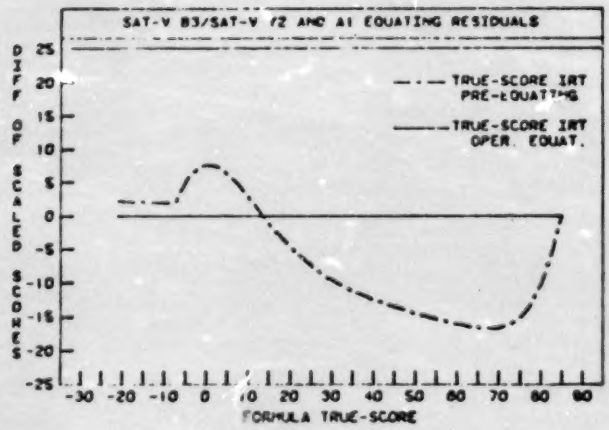


Figure 12: SAT-V Form 3ASA3 equated to SAT-V Form XSA2 and SAT-V Form YSA3 - Plots of raw score to scaled score transformations resulting from three IRT equatings compared to corresponding linear equating raw score to scaled score transformation.






SAT-V B3/SAT-V Y2 AND A1 EQUATING RESIDUALS
DIFF OF SCALED SCORES
FORMULA TRUE-SCORE
TRUE-SCORE IRT PRE-EQUATING
TRUE-SCORE IRT OPER. EQUAT.

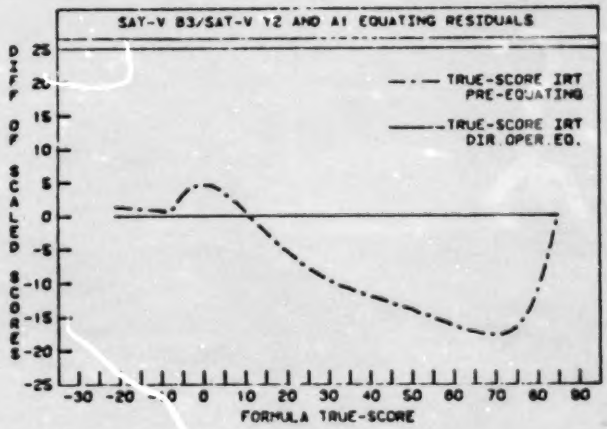


Figure 13: SAT-V Form 3BSA3 equated to SAT-V Form YSA2 and SAT-V Form 3ASA1 - Plots of 1) IRT pre-equating raw to scaled score transformation compared to corresponding IRT operational and direct operational equating raw score to scaled score transformations, and 2) Differences between scaled scores (IRT operational - IRT pre-equating) resulting from the equatings.

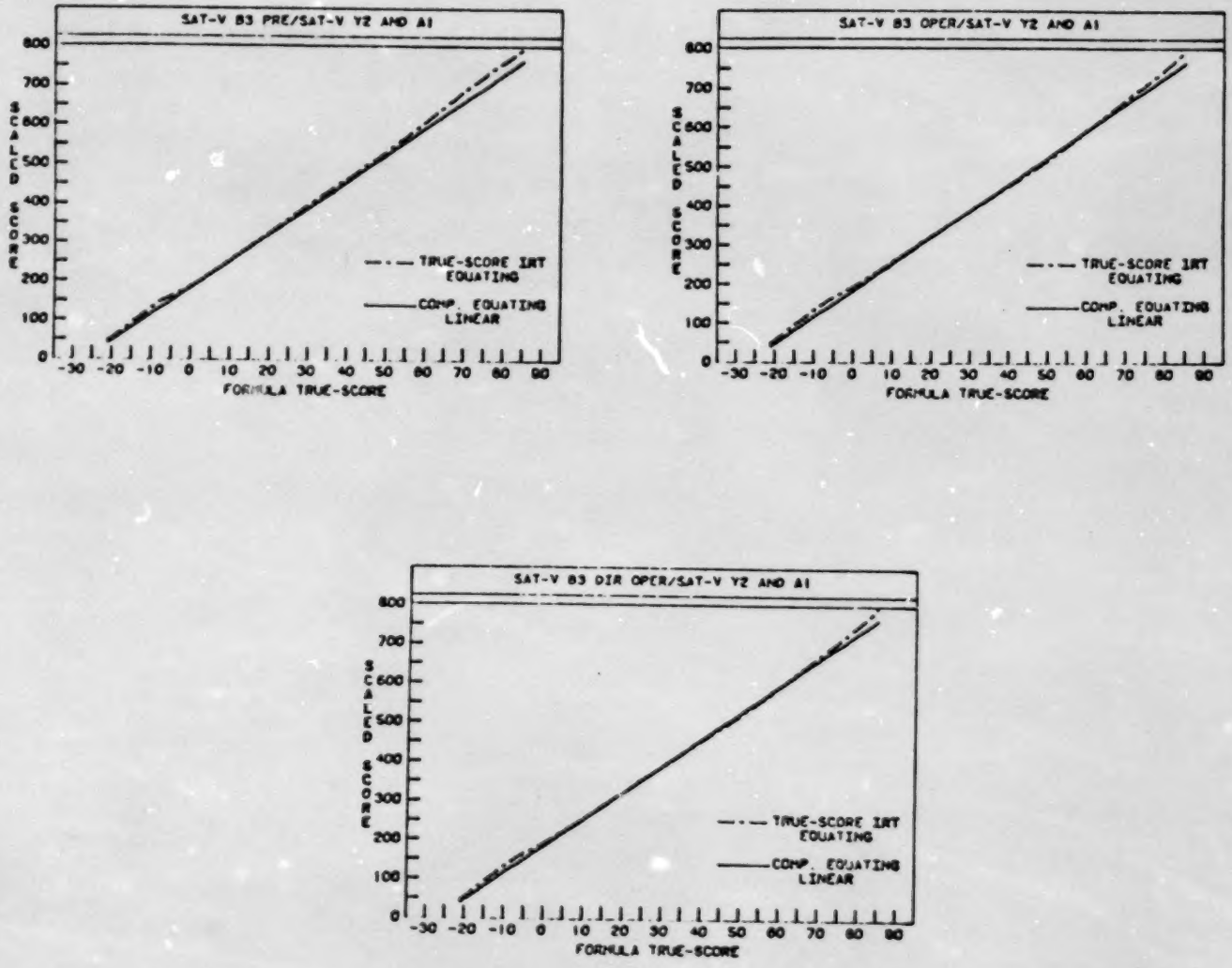


Figure 14: SAT-V Form 3BSA3 equated to SAT-V Form YSA2 and SAT-V Form 3ASA1 - Plots of raw score to scaled score transformations resulting from three IRT equatings compared to corresponding linear equating raw score to scaled score transformation.

Based on the data presented for Form 3ASA3, it is clear that the pre-equating was very successful. In no residual plot is the difference between the scaled score resulting from the pre-equating and the IRT operational or direct operational equating more than 15 score points. The pre-equating results in lower scaled scores at the bottom and top of the formula-score scale and slightly higher scaled scores in the middle region. As mentioned earlier, at no point are these differences greater than 15 scaled-score points, and hence, although the pattern of differences is consistent across equatings, the differences themselves are minor when compared to, for instance, the scaled standard error of measurement for SAT-Verbal which is approximately 30 scaled score points.

The pre-equating of Form 3BSA3 was not as successful as the pre-equating for Form 3ASA3 but, none-the-less quite adequate. The residual plots show differences in scaled scores resulting from the pre-equating and the operational or direct operational IRT equating of upwards of 20 scaled score points. The pre-equating results in slightly lower scaled scores at the lower end of the formula-score scale and consistently higher scaled scores through the middle and upper end of the formula-score scale. The maximum differences occur in all plots around a formula score of 70.

The results of pre-equating Form 3ASA3, when compared to the operational IRT equating, direct operational IRT equating, and linear equating, suggest that the pre-equating was very successful. The results of pre-equating 3BSA3 suggest that the pre-equating, while not as successful as for 3ASA3, was acceptable. Viewing both pre-equatings collectively in forming an overall evaluation of the study, suggests that the results are sufficiently promising to: 1) begin to consider pre-equating as a viable

method for equating SAT-Verbal; and 2) warrant the investigation of pre-equating for SAT-Mathematical.

IRT Equating with a Partial Pre-calibration Design

As mentioned previously, IRT equating for the SAT is currently based on item parameter estimates obtained from the concurrent calibration of items from the new edition, two equating tests, and two old editions of the test using data from two samples taking the new edition of the test and a sample from each group taking an old edition of the test. These item parameter estimates are all on the same scale and can be used without further adjustment for equating raw scores on the new edition to raw scores on each of the old editions. Since item parameter estimates from different administrations are not on a common scale, scores on a new test edition can only be equated to scores on old editions that were administered with the same equating sections as those given with the new edition.

In the future a more flexible procedure is envisioned to take advantage of item parameter estimates from different SAT administrations. Beginning with the 1983-84 testing year, it would be possible for the equating plans for the SAT to include item parameter equating sections to allow the equating of parameter estimates for items given in final test editions, equating tests, and pretests. This practice would mean that all future SAT items could be equated to the November 1982 base scale to be used in this study.

The study permits an evaluation of IRT equating that takes advantage of these new linkages, namely, equating based on partial pre-calibration. The essential feature of pre-calibration is that the items from the equating test are calibrated and placed on the base scale for the items prior to their administration with the new edition of the SAT. The pre-calibrated

items can be placed in an external equating test or an internal equating test (items embedded in the new edition). In practice, the external equating test would be the same as the equating tests now administered with the SAT. Score equating would take advantage of the fact that all item parameters for these equating sections would be on the same scale. This means that equating sections would be interchangeable, and any equating section could be used for equating purposes, not just those given with two particular old editions. The use of an internal equating test would be possible once enough pre-calibrated items were available for use in new editions. Thus, special equating tests would not be needed, allowing items formerly used in equating tests to go into the active item pool.

Both the external and internal cases provide flexibility with regard to the choice of which old editions to use in equating. The use of more old editions (only two are used currently) would presumably make scores more consistent from one administration to another. Since, in either case, items would be calibrated after the test administration, pre-equating would still not be possible. However, enough pre-calibrated items would presumably exist in the active item pool to permit the construction of some special editions (e.g., editions administered on Sunday) from pre-calibrated items, thus eliminating the need to spiral such editions in with other new editions and reducing the risk of a break in test security.

While this study will allow decisions to be made about the operational use of the partial pre-calibration design, it is limited to some extent in that the equating results will be compared to the results of operational equating rather than to a true criterion. The rationale for accepting the new design is based on the prior research that demonstrated the superiority

of IRT equating in SAT-like situations when a true criterion was available. However, further evaluation of the results of the partial pre-calibration design, particularly for the internal case and in a situation where the true equating results are known, will be made as soon as appropriate data are available. Once the partial pre-calibration design comes into operational use, such data will become available through the program.

In the study, item parameter estimates will be placed on a base scale through a number of concurrent calibrations and subsequent item parameter equatings of the results using common equating tests. Each of the concurrent calibrations will produce item parameters on a common scale; however, estimates from the separate calibrations will have to be equated. Among the equating tests to be calibrated in any given computer run in this study will be an equating test that is calibrated in another run. Thus, two sets of parameter estimates will exist for these items, and they will be on different scales because they will have resulted from two different calibration runs. Various procedures are available to obtain the linear relationship between the two sets of parameter estimates. This linear relationship is then used to place the item parameter estimates from the "current" run on the scale of the parameter estimates obtained in the other run. For the study an extensive design has been developed that will permit placing item parameter estimates for 25 SAT-Verbal editions and 25 SAT-Mathematical editions on the November 1982 item calibration scale. Items from equating sections and pretest sections that are calibrated along with the different editions of the SAT will also be placed on this scale.

Once the calibrations have been completed and all item parameter estimates have been placed on the November scale, it will be possible to equate the

scores from any particular edition of the SAT to the scores from any other edition of the SAT used in the study. In this study two types of score equating projected for future use will be simulated: score equating using a pre-calibrated external equating test (like those now given with the SAT) and score equating using a pre-calibrated internal equating test (items embedded in the operational test). For simulation purposes the operational test will be treated as if it had not been calibrated previously. The item parameters for the operational test will be placed on scale through two external equating tests or one internal equating test. IRT score equating will then be carried out to the two old editions used in the operational equating and also to a larger set of old editions.

Results of the different equatings will be compared to evaluate the effects of using these two types of score equating procedures. Because most of the data used in the study have been, or will be, obtained from the operational equatings of the SAT, an appropriate criterion exists - the operational equating resulting from the concurrent calibrations - against which to compare the results of the experimental equatings.

Summary

SAT-Verbal and SAT-Mathematical scores are now being equated by a method based on item response theory. In this paper, the authors have described the long range research and development program, preceeding the implementation of IRT equating, which was carried out to determine the desirability, the feasibility, and the probable effects of the change. The results of this extensive research and development effort showed that the new equating procedure will improve the comparability of scores on different editions of the SAT and the extent to which the SAT score scale maintains a constant meaning over time.

Current directions of research on IRT equating of the SAT were also discussed. These new directions take advantage of the flexibility of item response theory in that they examine the feasibility of partial or full pre-calibration and pre-equating of new editions of the test. The preliminary results of the pre-calibration and pre-equating studies appear to be favorable and it is anticipated that these techniques will be implemented to some degree for the SAT in the near future.

References

Angoff, W. H. Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, DC: American Council on Education, 1971.

Donlon, T. F., and Marco, G. L. Comparisons of the results of using linear and IRT equating methods to equate simulated SAT-verbal test forms assembled to new statistical specifications. Unpublished paper. Princeton, NJ: Educational Testing Service, 1981.

Eignor, D. R. An investigation of the feasibility and practical outcomes of pre-equating the SAT: Phase I interim report, verbal section. Unpublished paper. Princeton, NJ: Educational Testing Service, 1982.

Levine, R. S. Equating the score scales of alternate forms administered to samples of different ability (RB-55-23). Princeton, NJ: Educational Testing Service, 1955.

Lord, F. M. Estimation of latent ability and item parameters when there are omitted responses. Psychometrika, 1974, 39, 247-264.

Lord, F. M. Practical applications of item characteristic curve theory. Journal of Educational Measurement, 1977, 14, 117-138.

Lord, F. M. Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum, 1980.

Marco, G. L. Item characteristic curve solutions to three intractable testing problems. Journal of Educational Measurement, 1977, 14, 139-160.

Marco, G. L., Petersen, N. S., and Stewart, E. E. A test of the adequacy of curvilinear score equating models. Paper presented at the 1979 Computerized Adaptive Testing Conference, Minneapolis, 1979.

Petersen, N. S., and Cook, L. L. Feasibility of using IRT equating operationally for the SAT. Unpublished paper. Princeton, NJ: Educational Testing Service, 1981.

Petersen, N. S., Cook, L. L., and Stocking, M. S. Scale drift: A comparative study of IRT versus linear equating methods. Paper presented at the annual meeting of AERA, Los Angeles, 1981.

Petersen, N. S., Marco, G. L., and Stewart, E. E. An evaluation of linear and equipercentile equating methods. Paper presented at the ETS Research Conference on Test Equating, Princeton, NJ, April 2-3, 1980.

Petersen, N. S., and Whittington, C. D. The impact of item response theory equating on changes in the score distributions of selected colleges and high schools from 1980 and 1981 (November and December administrations). Unpublished paper. Princeton, NJ: Educational Testing Service, 1982.

Potthoff, R. F. Equating of grades or scores on the basis of a common battery of measurements. In P. R. Krishnaiah, Multivariate Analysis. Academic Press, 1966.

Stocking, M. L., and Lord, F. M. Developing a common metric in item response theory. RR-82-25-ONR. Princeton, NJ: Educational Testing Service, 1982.

Wood, R L, and Lord, F. M. A user's guide to LOGIST. RM-76-4. Princeton, NJ: Educational Testing Service, 1976.

Wood, R L, Wingersky, M. S., and Lord, F. M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters. Research Memorandum 76-6. Princeton, NJ: Educational Testing Service, 1976.